Politecnico di Milano

Facoltà di Ingegneria
Corso di Laurea in Ingegneria Elettronica

Laboratoire des Signaux et Systèmes

Groupe Problèmes Inverses
CNRS - Université de Paris-Sud - Supélec

Maximum Likelihood Estimation of Hidden Markov Model Parameters, with Application to Medical Image Segmentation

Relatore: Prof. Sergio Bittanti
Correlatore: Dr. Jérôme Idier

Tesi di Laurea di
Andrea Ridolfi
matr. 605731

Anno Accademico 1996-1997

ai miei cari,
in particolare a mio fratello

Acknowledgements

This work represents a *tesi di laurea*, the final projet to obtain an engineering degree from the *Politecnico di Milano* (Milan, Italy), the institution where I've accomplished most of my studies.

It has been carried out at the *Groupe Problèmes Inverses* of the *Laboratoire des Signaux et Systèmes - Université de Paris-Sud, CNRS, Supélec* (Gif-sur-Yvette, France) from September 1996 to July 1997.

It would have never been possible to realise it without the precious help of many persons. Their help has allowed me to satisfy administrative duties, to gain more knowledge and to fulfill my encouragement needs. I wish to thank all of them and I'll do it in the following, in their respective languages.

Desidero ringraziare in primo luogo il professor Sergio Bittanti, del dipartimento di *Elettronica e Informazione* del *Politecnico di Milano*, per i consigli e l'aiuto datomi nella mia formazione universitaria e per aver accettato di essere il relatore di questa tesi. Nonostante le difficoltá causate dalla distanza, mi ha seguito constantemente in questo lavoro con preziosi suggerimenti ed utilissime osservazioni.

Pour la deuxième fois, Guy Demoment, désormais "Big Big Chef", m'a accueilli dans son bateau, pardon, équipe. C'est avec une émotion sincère que j'exprime ma reconnaissance au "capitaine" du Groupe Problèmes Inverses, qui, *quand le bateau fait naufrage, crie: "Je suis le maître à bord, sauve qui peut ! Le vin et le pastis d'abord !"* (G. Brassens). Jérôme Idier m'a "mis sur les rails" et a suivi mon travail pendant tout son déroulement. Qu'il trouve ici l'expression de ma gratitude et mes excuses pour ne pas avoir répondu à ses attentes et pour l'avoir embêté avec mon mélange ital-fran-glais. Un merci très particulier à Gio. C'est avec plaisir et dévotion que je relis son fameux rapport interne. Heureux de son retour au laboratoire, j'espére qu'il m'expliquera comment régulariser pour allonger les AR. Je remercie également tout le GPI: I-AN pour s'être baladé avec moi dans le couloir, Papi Hervé pour m'avoir toujours sorti de mes problèmes informatiques, CuiCui pour ses bonbons, Steph B. et ses soupapes, Steph G. et ses essais d'italien, Dub pour la carte postale "classe", le pére Thierry et son Larousse vivant, P. Brémaud pour le culte du Général B., Berchos, les Ali, Champ et son verre dans le pied, et bien sur le Dieu de la Sauvegarde qui m'a protégé du quasi-hasard informatique.

Considerando che questo lavoro non è frutto solo di un anno di tesi ma soprattutto degli anni universitari che lo hanno preceduto, vorrei ringraziare tutti coloro che hanno fatto parte di questi anni: i professori che hanno saputo trasmettermi più del contenuto di un corso universitario e gli amici che hanno scambiato con me idee ed emozioni.

Fra i primi ringrazio in particolare, oltre al professor Sergio Bittanti, la professoressa Elisa Brinis Udeschini che ha trasmesso con anima le meraviglie del pensare Galileiano ed Einsteniano, il professor Domenico Pagani che mi ha seguito e consigliato durante i due anni di soggiorno all'estero nell'ambito di un programma di scambio europeo ed

il professor Giuseppe Drufuca per gli schietti consigli alle mie indecisioni dottorali.

Fra i secondi ringrazio in particolare "via Bassini 43" (Stefano per l'ospitalità nel suo castello e gli espropri di sua sorella, Ciccio per il trionfo commerciale, Gianluca per l'econo-filosofia serale, Gubi per le spremutine, Davide per le chiaccherate bianco-rosse e Ciano per romanzare *il barba, il presidente e la vespa*), "via Poggi" (Bobby for pope, Gio per l'alta fedeltà, Fabio per la somiglianza e Pier per la spiegazione del bonus), Cranio per le rivoluzioni dopo cena ed Alberto el il Bona per aver lavato i piatti (oltre che per il mal d'utopia), Giorgione per le chiaccherate notturne *in Spirit*, il Poz per i pacchi a pranzo, Valentina per gli appuntamenti culinari del lunedí ed Andrea per la valigetta rompispechietti.

Je tiens à remercier aussi M. Ing. Noubi Le Garant, pour avoir essayĆ de rendre crédible la double garantie "arabo-italienne", et Stephan Bogart et sa farine, avec laquelle on s'en n'est pas sorti. Un ringraziamento alla "grande comunità studentesca" italiana a Parigi, ed in particolare a Silvia per aver corretto l'estratto in italiano della tesi senza averci capito niente, a Lucia per il supporto morale ed a Stecco per il gelato alla nocciola da dieci franchi. A very special thank to Jennifer, for proof-reading my work during her ultimate all-you-can-see tour of Paris museums.

Contents

List of Figures

Notations

Notations for Spaces

$\mathbb{R}$: set of Real numbers
$\mathbb{R}^* = \mathbb{R} / \{0\}$
$\mathbb{R}_+$: set of positive Real numbers
$\mathbb{R}_+^* = \mathbb{R}_+ / \{0\}$
$\mathbb{N}$: set of Natural numbers
$\mathbb{N}^* = \mathbb{N} / \{0\}$
Θ : parameter space
$\overline{\Theta}$: closure of the parameter space Θ
$\mathcal{S}$: state space
$\mathcal{O}$: observation space

Notations for Vectors

$\boldsymbol{v}$: vector
v_k : element k of the vector $\boldsymbol{v}$
$\boldsymbol{v}_k^{k+l} = \{v_k \dots v_{k+l}\}$; $k, l \in \mathbb{N}^*$

Notations for Probabilities, Distributions and Stochastic Processes

$P(\boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{x})$: probability function
$f(\boldsymbol{X})$: density function for continuous random variables
$\boldsymbol{X} = \{X_k\}_{k \in \mathbb{N}^*}$: discrete time stochastic process
(k is the time index, and each X_k is a random variable)
$\boldsymbol{X} = \{X_{r,c}\}_{r,c \in \Lambda}$: stochastic process on a regular lattice Λ
(r, c are the spatial indexes of the lattice, and each $X_{r,c}$ is a random variable)
$\boldsymbol{X}_{r,1}^{r,C} = \{X_{r,c}\}_{r \in \{1,\dots,C\}}$: random vector by column
$\boldsymbol{X}_{1,c}^{R,c} = \{X_{r,c}\}_{r \in \{1,\dots,R\}}$: random vector by row
$\mathcal{G}_{m,\sigma^2}(y)$: Gaussian density function with variance σ^2 and mean value m

Notations for Indexes

$k,\ l$: time indexes
$i,\ j$: state indexes
$r,\ c$: spatial indexes of the regular lattice (row index and column index, respectively)
$k \in \{1, \dots, T\}$: k takes a value in $\{1, \dots, T\}$
$i = 1 \dots N$: i takes all the values from 1 to N

Abbreviations

EM : expectation maximization (algorithm)
HMC : hidden Markov chain
HMM : hidden Markov model
MAP : maximum *a posteriori*
MC : Markov chain
ML : maximum likelihood
MLE : maximum likelihood estimate
MRF : Markov random field
NLL : Negative log likelihood
PRF : Pickard random field
TEM : telegraphic EM (algorithm)
iid : independent and identically distributed (random variables)
e.g. : exempli gratia (for example)
i.e. : id est (that is)
$\triangle$: end of definition
$\square$: end of proof

Chapter 1

Introduction

1.1 Problem Overview

The present work deals with a study of the maximum likelihood estimation of hidden Markov model parameters, with an application to medical image segmentation.

The image we consider is issued from an X-ray tomography and it represents a slide view of the heart ventricle. It is a matrix of points (pixels), each one taking intensity values on a common space $\mathcal{O}$.

The calculation of the inner area of the ventricle, during the systolic and dystolic action, has a considerable medical interest. Due to measure blurring and noise it is difficult to clearly distinguish the surface of the ventricle from the surrounding membrane, and it is therefore difficult to compute the area. We then need an efficient method to distinguish the two surfaces by taking their nature into account: each one of them has a certain characteristic of uniformity.

For our purpose, we adopt the segmentation method, which consists of dividing the image in different zones according to a uniformity criterion that has to be defined. In other words, we can say that the segmentation operation "contrasts" the image: blurring is then eliminated while uniform areas and hedges are preserved.

For instance, let us consider a three level segmentation of the heart image. Such an operation may be described by the following two steps:

1. definition of the three levels;
2. division of the image in uniform zones (with at the most three different "types" of uniform zones), and level assignment to each zone (two zone have the same level if and only if they are of the same "type").

The three levels can be labeled to mark the zones: for instance, after the segmentation we may have uniform zones of type "A", of type "B" and of type "C". Alternatively, the levels may correspond to three values of intensity that are representative of the

original image: they can be some "weighted mean values" of the intensity distribution of the original image.

Intuitively, we can imagine that a three level segmentation of the heart image will assign the first level ("A") to the zone of the ventricle, the second level ("B") to the zone of the membrane and the last level ("C") to the zone outside the heart.

In practice, the division of the image is not as easy to perform as in the above example, and it requires an appropriate uniformity criterion. Such a criterion is implemented in the model we adopt in our study.

The model we consider is a *hidden Markov model* (HMM), that may be described as a double stochastic process. The first process models the segmented image with a Markov random field (MRF): this is the hidden process and has a Markovian characteristic. A second process models the observations: this is the observable process and is related to the hidden process through a certain law that has to be defined.

The Markovian character determines the uniformity criterion: it plays the role of an *a priori* information, modeling the notion of spatial uniformity while allowing sharp discontinuities. It is interesting to observe that, coherently with the "hidden" attribute of the process that models the segmented image, the latter is *a priori* unknown and the notion of spatial uniformity does not generate measurable quantities.

After the model structure has been defined, we aim to estimate the parameters of the model on the basis of the observation sequence. Such parameters are the parameters of the law that relates the observable process to the hidden one, and the parameters of the Markov random field. Only one parameter is imposed: the samples of the Markov random field are forced to take a finite number of values, and such a number is given by the desired number of segmentation levels.

Once the parameters are estimated, we can have access to the probability distribution of the Markov field. The segmented image is the most probable realization of the field.

The approach to the segmented image construction through the model parameter estimation is known as an *unsupervised problem*, in opposition to the *supervised problem* where the parameters are assigned *a priori*.

1.2 Background

Pioneer work on the subject is attributed to the authors of [Geman and Geman, 1984], although, their segmentation method is quite different from the one we use, since it is performed through contour identification instead of through level assignment. The hidden process is modeled as a general Markov random field and the most probable realization of the field is obtained by the maximization of the *a posteriori* probability distribution (that is the probability distribution of the field given the observations and the parameter estimates). Such a maximization is performed with a simulated

annealing technique.

The theoretical exhaustiveness and generality of their model is unfortunately affected by the slowness of the simulated annealing convergence [Woess, 1996]. Moreover, their work is limited to the supervised problem (*i.e.,* parameter are assigned *a priori*).

Despite the unidimensional case, the unsupervised problem for bidimensional hidden Markov models has not yet benefit from an exhaustive solution. Unidimensional hidden Markov models, which are also known as hidden Markov chains (HMCs), having a Markov chain as hidden process, have been successfully applied in the last twenty years to speech recognition. The related unsupervised problem has been deeply studied and a large set of parameter estimation techniques have been developed.

One may think to extend these techniques to bidimensional hidden Markov models, but unfortunately this is not possible, mainly due to the complex structure of Markov random fields with respect to Markov chains.

In order to overcome such a problem, the authors of [Devijver and Dekessel, 1988] propose to model the hidden process with a Pickard random field (PRF) [Pickard, 1980]. These are the only stationary symmetric Markov random fields of second-order on a finite lattice [Champagnat *et al.*, 1998]. The characteristic of the Pickard random fields is that each row and column is a reversible Markov chain.

According to [Devijver and Dekessel, 1988], the benefit of having a Markov chain on each row and column is to allow the use of hidden Markov chain techniques even in multidimensional contexts, at the expense of moderate approximations. Consequently, the unsupervised problem of a bidimensional hidden Markov model may be treated with a unidimensional approach by rows and columns.

To strengthen the theoretical background of the unidimensional approach to image segmentation proposed in [Devijver and Dekessel, 1988], in the recent original contribution [Goussard *et al.*, 1997] and in the previous work [Idier and Goussard, 1995], the Markov chain of each row and column is parameterized with telegraphic parameters [Godfrey *et al.*, 1980]. With such a parameterization the Markov chain is intrinsically reversible: this is an essential property to validate the unidimensional approach to bidimensional hidden Markov models based on Pickard random fields.

As discussed in [Goussard *et al.*, 1997], the telegraphic parameters fail to be easily estimate within the framework of the classical algorithm used for model parameter estimation (EM algorithm). To overcome such a problem, in [Goussard *et al.*, 1997] a new algorithm is proposed, based on a modification of the classical one.

The articles [Devijver and Dekessel, 1988] and [Goussard *et al.*, 1997] bring an extremely original contribution to the unsupervised segmentation problem, and more generically to the unsupervised problem for n-dimensional hidden Markov models. However, few problems are left unsolved, and in our work we try to give satisfactory answers to these problems.

1.3 Organization of the Work

The first part of this work exposes the theory of hidden Markov models.

In **Chapter 2**, we define the unidimensional models and we characterize their components: the observable process is supposed to be related to the hidden one by means of independent Gaussian distributions; the hidden process (the Markov chain) is parameterized with a standard set of parameters and then with a telegraphic set of parameters.

In **Chapter 3**, we refer to the maximum likelihood approach in order to estimate the model parameters. At this stage, we face the problem of the degeneracy of the likelihood function in the origin of the Gaussian distribution variance parameter. This is a well identified issue in such a mixture identification context [Nádas, 1983], but according to our knowledge it has not yet been clearly solved. After a precise analysis of the problem and a proof of the existence of a degeneracy point (Section 3.2), we study a Bayesian solution (Section 3.2.1) and analyze its properties. Such a study is the first original contribution of our work.

The maximization of the likelihood is performed with the algorithm known as *Expectation Maximization* (EM) [Baum *et al.*, 1970] (Section 3.4). It is a fixed point numerical method which maximizes the likelihood function through the maximization of an auxiliary function. In order to implement the Bayesian solution with respect to the degeneracy, we must develop a penalized version of such an algorithm.

With such an algorithm, we estimate the parameters of the law that relates the observable process to the hidden one, and the parameters of the Markov chain with standard parameterization. When the parameterization of the Markov chain is telegraphic, the EM algorithm cannot be efficiently applied.

In **Chapter 4**, we consider the telegraphic EM algorithm for the estimation of the telegraphic parameters, that has been proposed in [Goussard *et al.*, 1997]. We contribute with an original study of the auxiliary function of such an algorithm and we prove some interesting properties on the regular behavior of the function (Section 4.2). This study is the second original contribution of our work.

The properties we have found are unfortunately not sufficient to guarantee the systematic convergence of the telegraphic EM algorithm to a local maximum of the likelihood function. Moreover, in the case of a small set of observations, a counterexample has shown a situation in which the value of the likelihood oscillates during the algorithm iterations, and after a certain number of steps the algorithm stalls into a loop. It is then impossible to guarantee the systematic convergence of the algorithm to a local maximum of the likelihood function.

Nevertheless, with a large set of observations, the algorithm does not seem to suffer from such a problem.

Not being able to guarantee the local maximization of the likelihood function is a delicate problem. Additionally, the EM algorithm already suffers from a slow con-

vergence rate as discussed in [Campillo and Le Gland, 1989] and [Meilijson, 1989], and even when the local maximization is theoretically ensured, convergence time may be too high to attain the local maximum. This is clearly shown by the gradient of the likelihood: after a certain number of iterations the gradient stabilizes, but to a value different than zero (see Section 7.4).

However, even with such problems the EM algorithm still provides attractive features. In particular, the EM algorithm does not require augmentation with elaborate safeguards, such as those necessary for Newton's method and quasi-Newton methods, in order to produce iteration sequences with good global convergence characteristics.

The lack of a theoretical guarantee of convergence to a local maximum of the likelihood function and the problem of a slow convergence rate are overcome with a mixed EM - gradient descent technique. The performances of this mixed technique are analyzed in Section 7.4. This mixed method, that we still have to perfect, is the third original contribution of our work.

In **Chapter 5**, we introduce the bidimensional hidden Markov models based on Pickard random fields. The definition of the hidden process requires a particular attention, while the definition of the other components of the model may be extended from the unidimensional case. The model is expressly defined to fit our applied issue and to benefit from the unidimensional approach to parameter estimation.

In **Chapter 6**, we approach the model parameter estimation according to the article [Devijver and Dekessel, 1988], but in a telegraphic parameterization context as in [Goussard *et al.*, 1997]. Within the framework of a maximum likelihood technique, we adopt the approximated likelihood function proposed in [Devijver and Dekessel, 1988], which is the product of the marginal likelihood of each row and column of the observable process. By applying the EM algorithm principle to such an approximated likelihood function and by exploiting the Markovian property of the rows and columns of a Pickard random field, the model parameter estimates are obtained by means of an "accumulation" of the parameter estimates of each row and column.

In the second part of this work we expose our applied issue.

In **Chapter 7**, the bidimensional hidden Markov models are applied to the problem of medical image segmentation. In this chapter we justify the choices made in the definition of the model. The segmented image is modeled after the Pickard random field and, on the basis of the observations (the original image), we estimate an approximated form of the marginal *a posteriori* probability distribution of the random field, as suggested in [Devijver and Dekessel, 1988]. The segmented image is the realization of the random field that maximizes this approximated marginal *a posteriori* distribution.

The segmentation algorithms (parameter estimation algorithms and *a posteriori* maximization algorithms) are implemented in *Matlab* and *C* programming codes, and they are applied to the X-ray tomography image of the heart. In Section 7.4, we show and discuss the results of the segmentation.

Finally, we reach our conclusion and discuss the possible developments of our work.

Part I

Hidden Markov Models

Chapter 2

Unidimensional Hidden Markov Models

2.1 Introduction

Unidimensional hidden Markov models (HMMs) are doubly stochastic processes. They are composed by an underlying stochastic process $\boldsymbol{X} = \{X_k\}_{k\in\mathbb{N}^*}$ that is not observable (hidden) but can only be observed through another stochastic process $\boldsymbol{Y} = \{Y_k\}_{k\in\mathbb{N}^*}$ that produces the set of observations $\boldsymbol{y} = \{y_k\}_{k\in\mathbb{N}^*}$.

To understand the concept of the hidden Markov model (more precisely to understand what they can model) consider the following "coin tossing" example. Picture yourself in a room with a curtain through which you cannot see what is happening. On the other side of the curtain someone performs a single or multiple coin tossing experiment. You are told the results of the tossing but you have no idea how the tossing is performed. If you know that the tossing experiment is done with a single coin, it is easy to imagine how the results are generated: each result is the outcome of a coin flip. Nevertheless, if the tossing is performed with multiple coins, the situation is more complex. For instance, consider the tossing performed with three coins, one fair and two biased (In a biased coin tossing, head and tail appear with different probabilities, while in a fair coin tossing they appear with the same probability). The two biased coins are associated with the two faces of the fair coin, respectively. To report the outcome of every coin flip, the fair coin is flipped first in order to decide which biased coin to use, then the chosen biased coin is flipped and the result of this last toss is reported. It is clear that the results we obtain through the curtain are not sufficient to imagine how the experiment has been performed. Hence, the experiment is somehow "hidden".

Figure 2.1 provides a graphical representation of the above example, where we have supposed the biased coin #1 and the biased coin #2 to be associated to the tail face and the head face of the fair coin respectively. Tossing of the fair coin gives tail, while tossing of the biased coin #1 gives head: the result is then head.

The tossing experiment can be seen as the hidden process, and its hidden characteristic is clear if we consider the last example, while the communication of the results

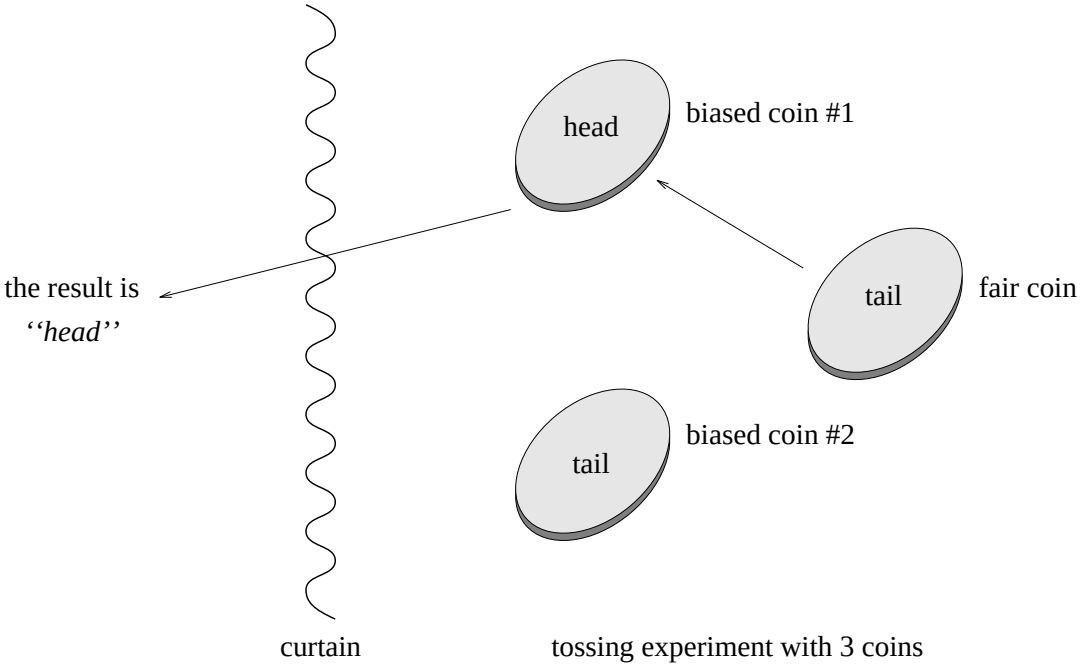

Figure 2.1: Tossing Experiment with 3 Coins

can be seen as the observable process that produces the set of observations. Thus, hidden Markov models apply to the modeling of such an example.

In order to define the hidden Markov model we need to specify the structure of the hidden process (the type of experiment *i.e.,* coin tossing) and the relation between the observable and the hidden process (what kind of information is given about the activities behind the curtain).

Let us first characterize the hidden process. We suppose that the hidden process is a homogeneous Markov chain:

Definition 2.1 Markov Chain
A Markov chain is a discrete time stochastic process $\boldsymbol{X} = \{X_k\}_{k \in \mathbb{N}^*}$. Each X_k ; $k \in \mathbb{N}^*$ is a random variable sampled on a common finite (numerable) state set $\mathcal{S}$: this means that the realizations $\boldsymbol{x} = \{x_k\}_{k \in \mathbb{N}^*}$ of the stochastic process take values on $\mathcal{S}$. Such values are called the *states* of the chain. Moreover, the stochastic process satisfies the following fundamental property:

$$P\left(\boldsymbol{X}_{k+1}^{k+l} \mid \boldsymbol{X}_1^k = \boldsymbol{x}_1^k\right) = P\left(\boldsymbol{X}_{k+1}^{k+l} \mid X_k = x_k\right) \ \ \forall k, l \in \mathbb{N}^* \tag{2.1}$$

This property is called the *Markov property of order one.*

If we consider the Markov chain over a time interval of length T (Markov chain of

length T), the probability of the realizations of the chain reads

$$P(\boldsymbol{X} = \boldsymbol{x}) = P(X_1 = x_1) \prod_{k=2}^{T} P(X_k = x_k \,|\, X_{k-1} = x_{k-1}) \tag{2.2}$$

We define

$$p_{x_1} = P(X_1 = x_1) \;\; : \;\; \textit{initial probabilities} \text{ of the chain} \tag{2.3}$$

where $x_1 \in \mathcal{S}$, and

$$P_{x_{k-1}x_k} = P(X_k = x_k \,|\, X_{k-1} = x_{k-1}) \;\; : \;\; \textit{transition probabilities} \text{ of the chain} \tag{2.4}$$

where $x_k,\ x_{k-1} \in \mathcal{S},\ k = 2 \ldots T$.

The initial and the transition probabilities characterize the Markov chain. The initial state of the chain (the realization of X_1, where 1 is the initial instant) is chosen according to the initial probabilities *i.e.,* the probability that $X_1 = x_1$ is given by p_{x_1}. Successively, say at the instant k, if the realization of X_{k-1} is x_{k-1}, the probability that the realization of X_k takes the value x_k is given by the transition probability $P_{x_{k-1}x_k}$: we say that if the chain is in a certain state (say at the instant $k-1$), it switches to another state (at the instant k) according to the transition probabilities. △

Definition 2.2 Homogeneous Markov Chain
A Markov chain (Definition 2.1) is said to be *homogeneous* if the transition probabilities (2.4) fulfill the homogeneity property

$$P(X_{k+1} = x_{k+1} \,|\, X_k = x_k) = P(X_{l+1} = x_{k+1} \,|\, X_l = x_k) \;;\; \forall k, l \in \mathbb{N}^*, \tag{2.5}$$

Roughly speaking, the above property means that the transition probabilities are time invariant.

Note that, from the property of homogeneity (2.5), the transition probabilities $P_{x_{k-1}x_k}$ of the homogeneous chain depend on the values assumed by x_{k-1} and x_k, but do not depend on the instant k. △

The Markov property of the hidden process gives the Markov attribute to the model. As a consequence of the assumption that the hidden process is a Markov chain, unidimensional hidden Markov models are often addressed as hidden Markov chains (HMCs) [Rabiner and Juang, 1986].

We suppose that the state space of the Markov chain has N elements (states) *i.e.,* $|\mathcal{S}| = N$, which are the values assumed by the realizations of the chain. We denote

these values with $m_1 \dots m_N$, thus $\mathcal{S} = \{m_1, \dots, m_N\}$. For notation ease we will often refer to m_i simply with its index i. In order to not create ambiguity, we name i as *state* and m_i as *state value*. Thus, if the realization x_k of the element X_k assumes a state $i \in \{1, \dots, N\}$ ie $x_k = i$, the corrisponding state value is given by m_{x_k}.

With the assumption $|\mathcal{S}| = N$, the homogeneous Markov chain depends on N initial probabilities and N^2 transitions probabilities. These quantities can be considered the parameters that characterize the chain. Alternatively, we can define the initial and the transition probabilities by means of another set of parameters, and consider the latter as the characterizing set of parameters of the chain. In any case, we denote such a set with $\boldsymbol{\theta}_X$, and in the following the notation will take the dependency of the chain on the parameters $\boldsymbol{\theta}_X$ into account.

The meaning of the transition probabilities has been discussed above in the definition of a Markov chain. In the case of a homogeneous Markov chain (our case) it is easy to provide a graphical representation of such a meaning, since the dependence on the time instant can be omitted. In figure 2.2 we have represented the *transition graph* of a homogeneous Markov chain with 3 possible states.

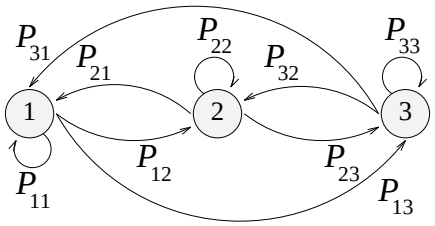

Figure 2.2: Transition Graph of an MC with 3 States

When the chain is in a certain state, i, at the successive instant it can remain in the same state with probability $P_{ii}(\boldsymbol{\theta}_X)$ or switch to a different state, j, with probability $P_{ij}(\boldsymbol{\theta}_X)$.

Figure 2.3 provides a timewise representation of the transition graph of a Markov chain with 3 states. Note that we have also represented the Markov chain at the initial instant, where the chain is in a certain state, i, with probability p_i.

Let us now characterize the observable process. The realizations $\boldsymbol{y} = \{y_k\}_{k \in \mathbb{N}^*}$ of the observable process $\boldsymbol{Y} = \{Y_k\}_{k \in \mathbb{N}^*}$ are supposed to take values on a continuous finite set $\mathcal{O} \subseteq \mathbb{R}$, and to be observed over a finite time interval *i.e.*, $\boldsymbol{y}_1^T = \{y_1 \dots y_T\}$ ($\boldsymbol{y}_1^T$ is denoted in the following simply as $\boldsymbol{y}$). Moreover, we assume that the observable process $\boldsymbol{Y} = \{Y_k\}_{k \in \mathbb{N}^*}$ is related to the hidden one $\boldsymbol{X} = \{X_k\}_{k \in \mathbb{N}^*}$ by N independent Gaussian distributions

$$f(Y_k \mid X_k = x_k) = \mathcal{G}_{m_{x_k}, \sigma^2_{x_k}}(Y_k) \; ; \; x_k \in \{1, \dots, N\}, \; k \in \mathbb{N}^* \tag{2.6}$$

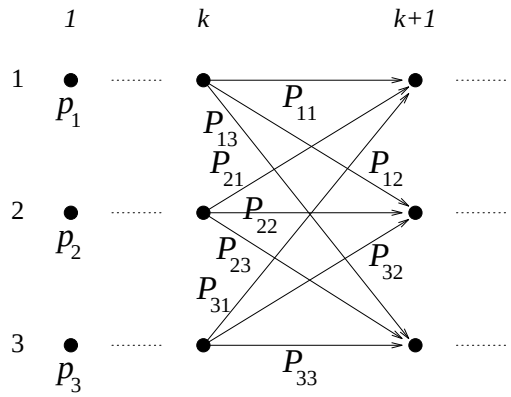

Figure 2.3: Timewise Representation of an MC with 3 States

where

$$\mathcal{G}_{m_{x_k},\sigma^2_{x_k}}(Y_k) = \frac{1}{\sqrt{2\pi\sigma^2_{x_k}}} \exp\left\{-\frac{(Y_k - m_{x_k})^2}{2\sigma^2_{x_k}}\right\} \tag{2.7}$$

is a Gaussian distribution with variance $\sigma^2_{x_k}$ and mean value m_{x_k} (note that the mean value of such a Gaussian distribution is the actual value of the realization x_k *i.e.,* is the state that corresponds to x_k).

From definition (2.6), the sequence $\boldsymbol{Y}_1^T = \{Y_1 \dots Y_T\}$ (denoted in the following simply as $\boldsymbol{Y}$), that generates the realizations over the observation lapse, is related to the hidden process by the joint distribution

$$f(\boldsymbol{Y} \mid \boldsymbol{x}) = \prod_{k=1}^{T} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}(Y_k) \tag{2.8}$$

This joint distribution is named the *observable characteristic* of the hidden Markov chain. It depends on $2N$ parameters denoted as $\boldsymbol{\theta}_{Y \mid X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$, where the parameters $\boldsymbol{m} = \{m_i \mid m_i \in \mathcal{S};\ i = 1 \dots N\}$ are the values of the states, and the parameters $\boldsymbol{\sigma}^2 = \left\{\sigma_i^2 \mid \sigma_i^2 \in \mathbb{R}_+^*;\ i = 1 \dots N\right\}$ are the variances of the Gaussian distributions.

In figure 2.4 a graphical representation of the hidden Markov model is provided. The hidden layer represents the N states $1 \dots N$ of the Markov chain (the hidden process) repeated at the time instants $1 \dots T$. We have considered a Markov chain that evolves from the state 1 at the initial instant 1 to the state 2 at the final instant T, through the states N, 2, 1 and 2, at the instants 2, 3, k and $T-1$, respectively. At each time instant, an observation y is produced on the observable layer (by the observable process), according to the current state value of the hidden process.

Together, the observable characteristic and the hidden process describe the hidden

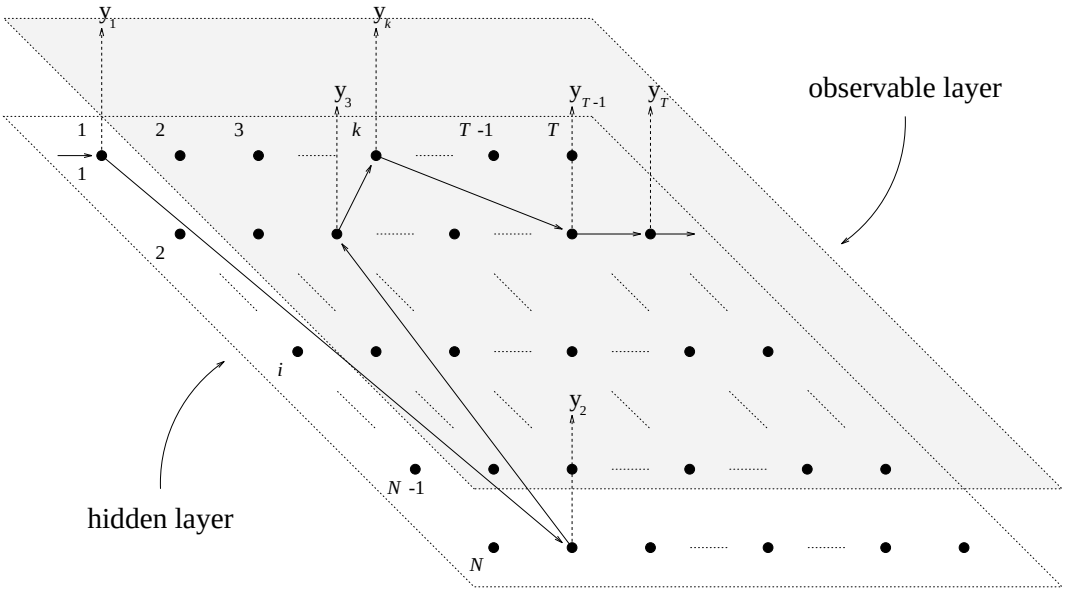

Figure 2.4: Graphical Representation of an HMC

Markov model, through the equations

$$f\left(\boldsymbol{Y} \mid \boldsymbol{x}\,;\,\boldsymbol{\theta}_{Y\mid X}\right) = \prod_{k=1}^{T} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}\left(Y_k\right)$$

$$P\left(\boldsymbol{X} = \boldsymbol{x}\,;\,\boldsymbol{\theta}_X\right) = p_{x_1}\left(\boldsymbol{\theta}_X\right) \prod_{k=2}^{T} P_{x_{k-1}x_k}\left(\boldsymbol{\theta}_X\right) \tag{2.9}$$

and the parameters $\boldsymbol{\theta}_{Y\mid X}$ and $\boldsymbol{\theta}_X$. We denote with $\boldsymbol{\theta} = \left\{\boldsymbol{\theta}_{Y\mid X}, \boldsymbol{\theta}_X\right\}$ the whole set of the hidden Markov chain parameters, and Θ as the space of their admissible values.

The first set of parameters has been already defined as $\boldsymbol{\theta}_{Y\mid X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$, while the second set depends on the parameterization chosen for the Markov chain. We consider two different parameterizations, as discussed in the following sections. Each one of them will give different properties to the hidden Markov model.

2.2 Standard Hidden Markov Chain

Let us define a *standard* hidden Markov chain as an hidden Markov model where the hidden process is a Markov chain (2.2) (supposedly homogeneous and with N states) parameterized with N initial probabilities (2.3) and the N^2 transition probabilities (2.4).

These parameters are subjected to the probability constraint

$$\sum_i p_i = 1\ ,\ p_i \geq 0\,;\ \ \forall i \in \{1 \ldots N\} \tag{2.10}$$

$$\sum_j P_{ij} = 1\ ,\ P_{ij} \geq 0\,;\ \ \forall i,j \in \{1 \ldots N\} \tag{2.11}$$

This is the most general homogeneous Markov chain, and we may refer to it as a *standard* Markov chain. Such a chain depends on $N + N^2$ parameters $\boldsymbol{\theta}_X = \{\boldsymbol{p}, \boldsymbol{P}\}$, where $\boldsymbol{p} = \{p_i\,;\ i = 1 \ldots N\}$ and $\boldsymbol{P} = \{P_{ij}\,;\ i,j = 1 \ldots N\}$.

In figure 2.5 we have represented a situation in which the standard Markov chain remains in the state i from the instants $k-1$ to the instant k, and then switches to the state j at the instant $k+1$. Part of the transition graph and the timewise representation of the chain are illustrated.

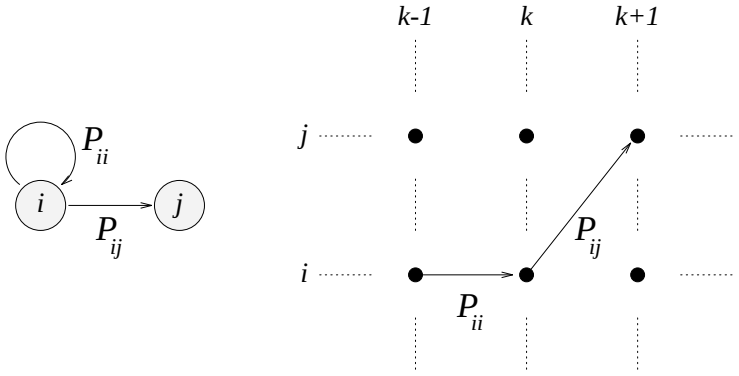

Figure 2.5: Transition Graph and Timewise Representation of a Standard MC

With a standard Markov chain as hidden process, the double stochastic model is described by the equations

$$f\left(\boldsymbol{Y} \mid \boldsymbol{x}\,;\, \boldsymbol{\theta}_{Y\mid X}\right) = \prod_{k=1}^{T} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}\left(Y_k\right)$$

$$P\left(\boldsymbol{X} = \boldsymbol{x}\,;\, \boldsymbol{\theta}_X\right) = p_{x_1} \prod_{k=2}^{T} P_{x_{k-1}x_k} \tag{2.12}$$

and the $N\,(3+N)$ parameters

$$\boldsymbol{\theta} = \left\{\boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{p}, \boldsymbol{P}\right\} \tag{2.13}$$

2.3 Telegraphic Hidden Markov Chain

Let us define a telegraphic hidden Markov chain as an hidden Markov model where the hidden process is a telegraphic Markov chain [Godfrey *et al.*, 1980] (supposed with N states).

Definition 2.3 Telegraphic Markov Chain
An N states telegraphic Markov chain is an N states homogeneous Markov chain with transition probabilities defined by mean of the parameters $\boldsymbol{\mu} = \{\mu_i \,;\, i = 1 \dots N\}$ and $\boldsymbol{\lambda} = \{\lambda_i \,;\, i = 1 \dots N\}$

$$P_{ij}\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) = \left(1 - \lambda_i\right) \mu_j + \delta_{ij} \lambda_i \,;\, i, j \in \{1 \dots N\} \tag{2.14}$$

δ_{ij} is the Kronecker symbol *i.e.,*

$$\delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i = j \end{cases}$$

△

As initial probabilities $(p_i\left(\boldsymbol{\theta}_X\right) \,;\, \in \{1, \dots, N\})$ we adopt the *stationary probabilities* of the chain $(\pi_i\left(\boldsymbol{\theta}_X\right) \,;\, \in \{1, \dots, N\})$ (see the definition hereafter), where the stationary probabilities of a telegraphic Markov chain are given by

$$\pi_i\left(\boldsymbol{\mu}, \boldsymbol{\lambda}\right) = \frac{\mu_i / \left(1 - \lambda_i\right)}{\sum\limits_j \mu_j / \left(1 - \lambda_j\right)} \,;\, i \in \{1, \dots, N\} \tag{2.15}$$

Definition 2.4 Stationary Probability of a Homogeneous Markov Chain
Let us consider an N states homogeneous Markov chain, with transition probabilities $P_{ij}\left(\boldsymbol{\theta}_X\right) \,;\, i, j \in \{1, \dots, N\}$. The probabilities $\pi_i\left(\boldsymbol{\theta}_X\right) \,;\, i \in \{1, \dots, N\}$ are said to be the *stationary probabilities* of the chain if and only if

$$\sum_i \pi_i\left(\boldsymbol{\theta}_X\right) P_{ij}\left(\boldsymbol{\theta}_X\right) = \pi_j\left(\boldsymbol{\theta}_X\right) \;\; \forall j \in \{1, \dots, N\} \tag{2.16}$$

△

The initial and transition probabilities are now both defined by means of the parameters $\{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$. The latter are then adopted as the characterizing parameters of the chain *i.e.,* $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$. Hence, the telegraphic Markov chain depends only on $2N$ parameters.

These parameters must satisfy the probability constraints on the initial and transition probabilities

$$\sum_i p_i(\boldsymbol{\mu}, \boldsymbol{\lambda}) = 1 \qquad \textit{(intrinsically satisfied)}$$

$$p_i(\boldsymbol{\mu}, \boldsymbol{\lambda}) \geq 0 \Leftrightarrow \frac{\mu_i/(1-\lambda_i)}{\sum_j \mu_j/(1-\lambda_j)} \geq 0\,; \;\; \forall i \in \{1, \ldots, N\} \tag{2.17}$$

$$\sum_j P_{ij}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = 1 \Leftrightarrow (1-\lambda_i)\sum_j \mu_j + \lambda_i = 1\,; \;\; \forall i \in \{1, \ldots, N\} \tag{2.18}$$

$$P_{ij}(\boldsymbol{\mu}, \boldsymbol{\lambda}) \geq 0 \Leftrightarrow (1-\lambda_i)\,\mu_j + \lambda_i \geq 0\,; \;\; \forall i, j \in \{1, \ldots, N\} \tag{2.19}$$

As a consequence of the choice of the stationary probabilities as initial probabilities, the telegraphic Markov chain is a reversible chain.

Definition 2.5 Reversible Homogeneous Markov Chain
A homogeneous Markov chain (with N states) with initial probabilities $p_i(\boldsymbol{\theta}_X)$; $i \in \{1, \ldots, N\}$ and transition probabilities $P_{ij}(\boldsymbol{\theta}_X)$; $i, j \in \{1, \ldots, N\}$ is said to be reversible if and only if

$$P_{ij}(\boldsymbol{\theta}_X)\, p_i(\boldsymbol{\theta}_X) = P_{ji}(\boldsymbol{\theta}_X)\, p_j(\boldsymbol{\theta}_X)\;; \;\; \forall i, j \in \{1, \ldots, N\} \tag{2.20}$$

△

Property 2.1 *A telegraphic Markov chain with stationary probabilities as initial probabilities is a reversible chain.*

Proof 2.1 Let us consider the transition probabilities (2.14) and the stationary probabilities (2.15) as initial probabilities. If we exclude the trivial case $i = j$, in which the reversibility condition (2.20) is automatically satisfied, we obtain

$$P_{ij}(\boldsymbol{\mu}, \boldsymbol{\lambda})\, p_i(\boldsymbol{\mu}, \boldsymbol{\lambda}) = (1-\lambda_i)\,\mu_j \frac{\mu_i/(1-\lambda_i)}{\sum_j \mu_j/(1-\lambda_j)}$$

$$= \frac{\mu_i \mu_j}{\sum_j \mu_j/(1-\lambda_j)} = P_{ji}(\boldsymbol{\mu}, \boldsymbol{\lambda})\, p_j(\boldsymbol{\mu}, \boldsymbol{\lambda}) \tag{2.21}$$

□

Finally we can consider the telegraphic hidden Markov chain described by the equations

$$f\left(\boldsymbol{Y} \mid \boldsymbol{x}\,;\, \boldsymbol{\theta}_{Y\mid X}\right) = \prod_{k=1}^{T} \mathcal{N}\left(y_k - m_{x_k}, \sigma_{x_k}\right)$$
$$P\left(\boldsymbol{X} = \boldsymbol{x}\,;\, \boldsymbol{\theta}_X\right) = \frac{\mu_{x_1}/\left(1-\lambda_{x_1}\right)}{\sum_i \mu_i/\left(1-\lambda_i\right)} \prod_{k=2}^{T} \left\{\left(1-\lambda_{x_{k-1}}\right)\mu_{x_k} + \delta_{x_{k-1}x_k}\lambda_{x_{k-1}}\right\} \tag{2.22}$$

and the $4N$ parameters

$$\begin{aligned} \boldsymbol{\theta}_X &= \{\boldsymbol{\mu}, \boldsymbol{\lambda}\} \\ \boldsymbol{\theta}_{Y\mid X} &= \left\{\boldsymbol{m}, \boldsymbol{\sigma}^2\right\} \end{aligned} \tag{2.23}$$

Note that, with respect to the standard hidden Markov chain, the telegraphic hidden Markov chain has a reduced number of model parameters. Moreover, the telegraphic Markov chain benefits from the reversibility property (Property 2.1). The latter achievement will be fundamental in validating our unidimensional approach to image segmentation (see Chapter 5 and Chapter 6).

2.3.1 Interpretation of the Telegraphic Parameterization

The telegraphic parameterization $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ can be interpreted as follows. The transition from one state to another is the result of a "double tossing" experiment. On the basis of the first toss, the decision of whether or not change the actual state i is made

$$\textit{first toss:} \begin{cases} \textit{we change the } i \textit{ state with probability } 1-\lambda_i \\ \textit{we keep the } i \textit{ state with probability } \lambda_i \end{cases}$$

In case the decision of state change is made, a second toss is performed in order to select the new state:

$$\textit{second toss: the } j \textit{ state is chosen with probability } \mu_j$$

In conclusion, the transition from state i to a different state j is performed with a probability $(1-\lambda_i)\,\mu_j$, while the transition from state i to itself results from the following procedure: one can either decide to keep state i (with probability λ_i), or decide to change state i (with probability $1-\lambda_i$) and then choose state i as the new

state (with probability μ_i). This mean that the transition from state i to itself is performed with a probability $\lambda_i + (1 - \lambda_i)\,\mu_i$.

Figure 2.6 shows, on the left, part of the transition graph of the chain, within the "double tossing" interpretation framework. The two different ways of performing the transition from state i to itself should be now more clear. In the right we give a timewise representation of the chain over three time instants. As an example, we have considered the transition of the chain from state i (at time instant $k-1$) to state j (at time instant $k+1$), through state i (at time instant k).

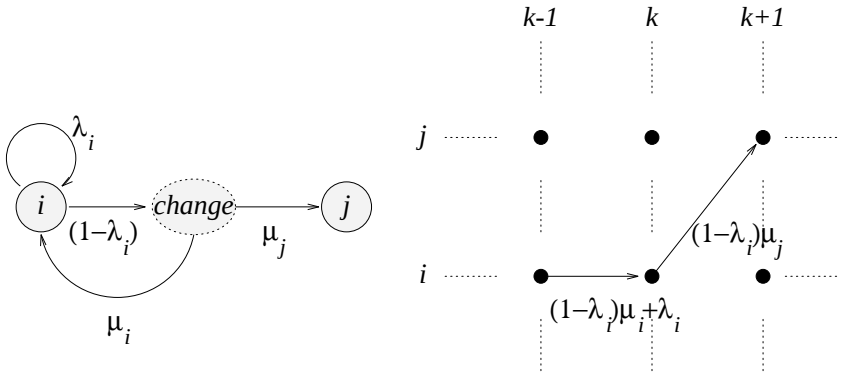

Figure 2.6: "Double Tossing" Interpretation of a Telegraphic MC

With such an interpretation, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are both probabilities; therefore, they are both submitted to the probability constraints

$$\sum_i \mu_i = 1\ ,\ \ \mu_i \geq 0\,;\ \ \forall i \in \{1 \ldots N\} \tag{2.24}$$

$$\sum_i \lambda_i = 1\ ,\ \ \lambda_i \geq 0\,;\ \ \forall i \in \{1 \ldots N\} \tag{2.25}$$

We must remark that in order to fulfill the constraints on the initial and transition probabilities (2.17-2.19), it is sufficient to consider the constraint (2.24) together with

$$-\frac{\mu_i}{1-\mu_i} \leq \lambda_i \leq 1 \tag{2.26}$$

It is then excessive to constrain both the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ as probabilities. If we consider that parameters estimation is more complex when the parameters are submitted to constraints, the "double tossing" interpretation of the transition of the chain reveals itself to be quite restrictive.

We can interpret the transition from one state to another simply as the result of a "single tossing" experiment (with transition probabilities given by (2.14)): the $\boldsymbol{\lambda}$ param-

eters are then released from the probability meaning (and relative constraints (2.25)).

According to this last interpretation, part of the transition graph of the chain may be represented as in figure 2.7 (on the left). The example in the timewise representation (on the right) is the same as in figure 2.6. Note that such a representation does not change with respect to the "single tossing" case: the two interpretations of the transition from one state to another influence how the transition is conceived, but do not influence the value of the transition probabilities.

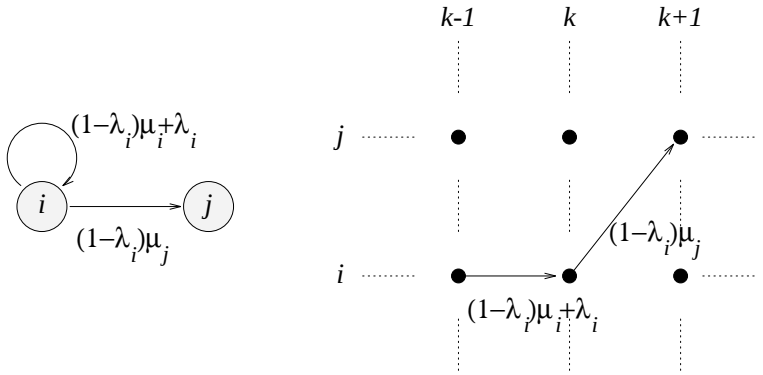

Figure 2.7: "Single Tossing" Interpretation of a Telegraphic MC

Chapter 3

Maximum Likelihood for HMC Parameter Estimation

3.1 Introduction

In this chapter we consider the problem of estimating the parameters of the hidden Markov chain (introduced in Chapter 2) on the basis of the observation sequence (the realizations of the observable process). For this purpose, the model parameters $\boldsymbol{\theta}_{Y\,|\,X}$ and $\boldsymbol{\theta}_X$ are adjusted to maximize the probability of the observation sequence given the model. More precisely, we select a parameter set which maximizes the *a posteriori* likelihood function. Such a choice is known as the *maximum likelihood estimate* (MLE) [Bittanti, 1993].

The likelihood function of our hidden Markov chain (2.9) is given by

$$\begin{aligned} f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) &= \sum_{\boldsymbol{x}} f\left(\boldsymbol{y}\,|\,\boldsymbol{x}\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right) P\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X\right) \\ &= \sum_{\boldsymbol{x}} P\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X\right) \prod_{k=1}^{T} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}\left(y_k\right) \end{aligned} \tag{3.1}$$

The maximization of the likelihood function is a difficult problem. As pointed out [Rabiner and Juang, 1986], there is no known way to find a solution in closed form. Moreover, the function degenerates in the origin of any of the parameters $\boldsymbol{\sigma}^2$: this is intuitively due to the presence of the parameters $\boldsymbol{\sigma}^2$ in the denominator of the Gaussian distributions.

In the following section we discuss the degeneracy and we propose a Bayesian solution. Then, in order to compute the maximum of the likelihood function, we face the problem of the evaluation of the function, and finally we approach its maximization with an iterative procedure: the *Expectation Maximization* (EM) algorithm.

3.2 Degeneracy of the Likelihood Function

The modeling of the data as continuous independent Gaussian random variables naturally induces a degeneracy toward infinity of the likelihood function. Intuitively, this is due to the fact that in a Gaussian distribution the variance parameter appears in the denominator. Therefore, the origin of such a parameter may cause the distribution, and the relative likelihood function, to degenerate.

This is a well known problem for mixture of Gaussian variables, as discussed in the paper [Redner and Walker, 1984], and in general for continuous unimodal densities [Nádas, 1983]. A different choice of distribution may not lead to degeneracy of the likelihood function. For instance, in [Leroux and Puterman, 1992] the observable process is related to the hidden one by means of a discrete distribution.

In our model (2.12) we have supposed the observations to be related to the hidden Markov chain through independent Gaussian distributions, hence we fall within the framework of the problem discussed above. More precisely, it can be observed that in the likelihood function of our model (3.1), the parameters $\boldsymbol{\sigma}^2 = \{\sigma_i^2, i = 1 \dots N\}$ appear in the denominator (through the Gaussian distributions $\mathcal{G}_{m_{x_k}, \sigma^2_{x_k}}(Y_k)$, $k = 1 \dots T$). Therefore, the origin of any $\sigma^2_{x_k}$, $x_k \in \{1, \dots, N\}$, may be a point of degeneracy of the likelihood function (this is the only possible point of degeneracy in the parameter space).

Indeed, we have proven the following property:

Property 3.1 *Let us consider the likelihood function (3.1), then*

$$\forall \boldsymbol{y} \in \mathcal{O},\ \exists\ \boldsymbol{\theta}^0 \in \bar{\Theta} \mid \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^0} f(\boldsymbol{y}; \boldsymbol{\theta}) = +\infty$$

where Θ is the parameter space, $\bar{\Theta}$ is the closure of such a space, $\boldsymbol{\theta}^0 = \left\{\boldsymbol{m}, \boldsymbol{\sigma}^{2^0}, \boldsymbol{\theta}_X\right\} \in \bar{\Theta}$ is a point in the closure of the parameter space, and $\boldsymbol{\theta} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}_X\} \in \Theta$ is a point in the parameter space.

Proof 3.1 For a fixed time l, let us choose the value of a given state n so to coincide

with the observation $(m_n = y_l)$ and let ${\sigma^2}^0_n = 0$ and ${\sigma^2}^0_i > 0 \; \forall i \neq n$. Then

$$\begin{aligned} f(\boldsymbol{y}\,;\,\boldsymbol{\theta}) &= \sum_{\boldsymbol{x}} P(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X)\, \mathcal{G}_{m_{x_l},\sigma^2_{x_l}}(m_n) \prod_{k\neq l} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}(y_k) \\ &\geq \sum_{\boldsymbol{x}\,|\,x_l=n} P(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X)\, \mathcal{G}_{m_n,\sigma^2_{x_l}}(m_n) \prod_{k\neq l} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}(y_k) \\ &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \sum_{\boldsymbol{x}\,|\,x_l=n} P(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X) \prod_{k\neq l} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}(y_k) \\ &\geq \frac{1}{\sqrt{2\pi\sigma_n^2}} \sum_{\boldsymbol{x}\,|\,\substack{x_i\neq n\forall i\neq l \\ x_l=n}} P(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X) \prod_{k\neq l} \mathcal{G}_{m_{x_k},\sigma^2_{x_k}}(y_k) \end{aligned} \tag{3.2}$$

In the last expression of (3.2), the product of the Gaussian terms excludes $k = l$, therefore it is affected only by the summation over $x_i \neq n \; \forall i \neq l$ and none of the σ_{x_k} realizations in the product attain the bound value zero. Under these conditions the Gaussian term does not attain zero in any finite support, and since the set of observations is finite with finite values, the product remains strictly positive. Moreover, let us consider any $\boldsymbol{\theta}_X$ so that the Markov chain fulfills the "positivity condition" *i.e.*, $\forall \boldsymbol{x},\; P(\boldsymbol{x}\,;\,\boldsymbol{\theta}_X) > 0$[1].

Finally we obtain

$$f(\boldsymbol{y}\,;\,\boldsymbol{\theta}) \geq \frac{K(\boldsymbol{\theta})}{\sqrt{\sigma_n^2}}\,, \quad \text{with } K(\boldsymbol{\theta}) = K\left(\boldsymbol{\theta}^0\right) > 0$$

As $\sigma_n^2 \to {\sigma_n^2}^0$ the right term tends to infinity and it ensures that $f(\boldsymbol{y}\,;\,\boldsymbol{\theta}) \to +\infty$.

$\boldsymbol{\sigma}^{2^0} = \left\{ {\sigma^2}^0_i \,|\, {\sigma^2}^0_n = 0, {\sigma^2}^0_i \neq 0 \; \forall i \neq n \right\}$ and $\boldsymbol{m} = \{m_i \,|\, m_n = y_l\}$ as well as the parameters $\boldsymbol{\theta}_X$ chosen so to guarantee the "positivity condition", constitute the point $\boldsymbol{\theta}^0 \in \bar{\Theta}$, in the closure of the parameters space, which causes the degeneracy of the likelihood function. □

The likelihood function is then not bounded above and its degeneracy point in $\bar{\Theta}$ is an attracting domain for the maximization of the function.

[1] The "positivity condition" is a restrictive constraint. However, since we search for sufficient conditions in order to prove the degeneracy of the likelihood function, the choice of the parameters is completely arbitrary and they can be submitted to any constraint. Nevertheless, it can be proved that the summation term is ensured to be strictly positive with a larger subset of parameters, such as a subset which guarantees the irreducibility of the Markov chain.

3.2.1 Bayesian Solution to Singularity

A Bayesian solution is proposed to solve the degeneracy of the likelihood function in $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^{2^0}$. The parameters $\boldsymbol{\sigma}^2$ are considered as random variables, while the other parameters $\{\boldsymbol{\mu}, \boldsymbol{\theta}_X\}$, are still considered as deterministic quantities. The likelihood function is then replaced by the *penalized* likelihood function

$$f\left(\boldsymbol{y}, \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) = f\left(\boldsymbol{y} \mid \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) f\left(\boldsymbol{\sigma}^2\right)$$

where $f\left(\boldsymbol{y} \mid \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right)$ is the likelihood function and $f\left(\boldsymbol{\sigma}^2\right)$ is the *a priori* distribution of the parameters $\boldsymbol{\sigma}^2$.

Intuitively, in order to eliminate the degeneracy, we need an *a priori* distribution on $\boldsymbol{\sigma}^2$ which, in a contour of the origin of any of the parameters $\boldsymbol{\sigma}^2$, tends to zero with at least the same order at which the likelihood degenerates (*i.e.,* as the likelihood tends to infinity).

Let us suppose the parameters $\boldsymbol{\sigma}^2$ to be *iid* random variables

$$f\left(\boldsymbol{\sigma}^2\right) = \prod_{i=1}^{N} f\left(\sigma_i^2\right) \tag{3.3}$$

distributed with a generic *a priori* exponential probability distribution

$$f\left(\sigma_i^2\right) = K_{norm} \frac{1}{\sigma_i^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma_i^{2\gamma}}\right\} 1_{[0,+\infty)} \tag{3.4}$$

The latter fulfills the intuitive property in order to eliminate the degeneracy. Further specification of its parameters must take the requirement of a proper distribution into account.

As will be discussed in Section 3.5.1, to maintain an explicit re-estimation formula for the parameters $\boldsymbol{\sigma}^2$ in the penalized EM algorithm, the parameter γ must be equal to one. This constraint leads, from the generic exponential *a priori* distribution (3.4), to an inverted gamma distribution

$$f\left(\sigma_i^2\right) = \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)} \frac{1}{\sigma_i^{2\beta}} exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} 1_{[0,+\infty)} \tag{3.5}$$

This distribution in ensured to be proper if

$$\beta > 1 \text{ and } \alpha > 0 \tag{3.6}$$

as discussed in [Robert, 1992].

The inverted gamma distribution is a classical choice for the *a priori* distribution of the variance of a Gaussian density, as pointed out in [Diebolt and Robert, 1994] and in [Robert, 1992]. Such a distribution is obtained when the only *a priori* information available is that the variance is a "scale" parameter [Demoment, 1996]. Within this classical framework, in [Besag *et al.*, 1991] the authors propose an *a priori* distribution similar to the inverted gamma that solves the degeneracy problem.

It is interesting to remark that the way we have obtained the inverted gamma distribution, as the *a priori* distribution for the variance parameter, is completely different from the classical one. Indeed, we have firstly required the *a priori* distribution to tend to zero with at least the same order as the likelihood tends to infinity, and then we have required the penalized version of the likelihood to give explicit re-estimation formulas (within the EM algorithm framework).

3.2.1.1 The Inverted Gamma Distribution

Let us consider the inverted gamma distribution. In figure 3.1 the distribution is drawn for set of parameters α. Note that, since the represented gamma distributions

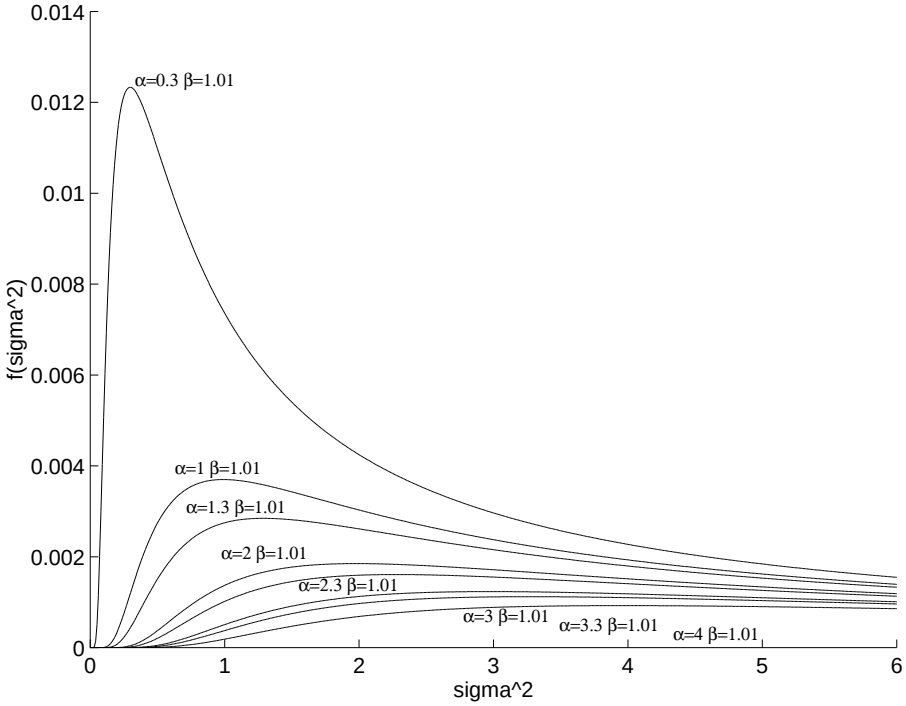

Figure 3.1: Inverted Gamma Distribution

are normalized, the tails of the curves must cross each other after a certain value of σ^2

(the area below each curve must be equal to one). Due to the limited horizontal scale of the figure the crossing of the curves cannot be seen.

The inverted gamma distribution benefits from the following property, which will be useful in proving the boundedness of a penalized version of the likelihood function:

Property 3.2 *The inverted gamma is a unimodal distribution.*

Proof 3.2 The bound values of the function are

$$\lim_{\sigma_i^2 \to 0^+} f\left(\sigma_i^2\right) = 0; \quad \lim_{\sigma_i^2 \to +\infty} f\left(\sigma_i^2\right) = 0$$

and since the function is continuous and positive over $(0, +\infty)$

$$\exists \max_{\sigma_i^2 \in (0,+\infty)} f\left(\sigma_i^2\right)$$

The maximum points of the function are given by the equation

$$\frac{\partial f\left(\sigma_i^2\right)}{\partial \sigma_i^2} = \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)\sigma_i^{2\beta}} \left(-\frac{\beta}{\sigma_i^2} + \frac{\alpha}{\sigma_i^4}\right) \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} = 0$$

which has only one solution in $(0, +\infty)$: $\sigma_i^2 = \dfrac{\alpha}{\beta}$. Hence, the maximum value is

$$\max_{\sigma_i^2 \in (0,+\infty)} f\left(\sigma_i^2\right) = \frac{\beta^\beta}{\Gamma\left(\beta-1\right)\alpha} \exp\left\{-\beta\right\}$$

Note that, as a function of α, the maximum value describes a hyperbola. In figure 3.1, where the inverted gamma is drawn for a set of parameters α, we may observe the hyperbola described by the maximum values. □

3.2.2 Penalized Likelihood Function

Let us denote the set of parameters $\boldsymbol{\theta}$ explicitly with $\{\boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}_X\}$, in order to consider separately the parameters $\boldsymbol{\sigma}^2$ ($\boldsymbol{\sigma}^2$ is random variable while the other parameters are deterministic quantities).

The penalized version of the likelihood function is given by

$$f\left(\boldsymbol{y}, \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) = f\left(\boldsymbol{y} \mid \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) f\left(\boldsymbol{\sigma}^2\right) \tag{3.7}$$

As a consequence of the likelihood expression (3.1) and the choice of independent *a priori* inverted gamma distributions (3.3) and (3.5), the penalized likelihood reads

$$\begin{aligned} f\left(\boldsymbol{y}, \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) &= f\left(\boldsymbol{y} \mid \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) f\left(\boldsymbol{\sigma}^2\right) \\ &= \sum_{\boldsymbol{x}} P\left(\boldsymbol{x} ; \boldsymbol{\theta}_X\right) \prod_{k=1}^{T} \frac{1}{\sqrt{2\pi\sigma_{x_k}^2}} \exp\left\{-\frac{\left(y_k - m_{x_k}\right)^2}{2\sigma_{x_k}^2}\right\} \\ &\qquad \prod_{i=1}^{N} \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)} \frac{1}{\sigma_i^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} \end{aligned} \tag{3.8}$$

Let us note with

$$\eta_i\left(\boldsymbol{x}\right) = \sum_{k=1}^{T} \delta_{ix_k}$$

the number of realizations assuming value i, where δ_{ix_k} is the Kronecker symbol, *i.e.,*

$$\delta_{ix_k} = \begin{cases} 0 \text{ if } i \neq x_k \\ 1 \text{ if } i = x_k \end{cases}$$

Then (3.8) may be written as

$$\begin{aligned} &\sum_{\boldsymbol{x}} P\left(\boldsymbol{x} ; \boldsymbol{\theta}_X\right) \frac{1}{\sqrt{2\pi}^T} \prod_{i=1}^{N} \frac{1}{\sigma_i^{\eta_i(\boldsymbol{x})}} \exp\left\{-\frac{\sum_{k=1}^{T} \delta_{ix_k} \left(y_k - m_i\right)^2}{2\sigma_i^2}\right\} \\ &\qquad \prod_{i=1}^{N} \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)} \frac{1}{\sigma_i^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} \\ &= \sum_{\boldsymbol{x}} P\left(\boldsymbol{x} ; \boldsymbol{\theta}_X\right) \frac{1}{\sqrt{2\pi}^T} \\ &\qquad \prod_{i=1}^{N} \frac{1}{\sigma_i^{\eta_i(\boldsymbol{x})+2\beta}} \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)} \exp\left\{-\frac{\sum_{k=1}^{T} \delta_{ix_k} \left(y_k - m_i\right)^2 + 2\alpha}{2\sigma_i^2}\right\} \end{aligned} \tag{3.9}$$

The following properties prove that the penalized likelihood does not degenerate any more at the origin of any of the parameters $\boldsymbol{\sigma}^2$, and that the origin of any of the parameters $\boldsymbol{\sigma}^2$ cannot be a maximum likelihood estimate *i.e.,* the likelihood function is not maximized on the lower extremity of the domain of the parameters $\boldsymbol{\sigma}^2$.

Property 3.3 *The penalized likelihood is bounded above over the parameters space (the penalized likelihood function does not degenerate on any point of the closure of*

parameters space $\overline{\Theta}$*).*

Proof 3.3 Akin to the likelihood function, the penalized version (3.8) may degenerate only in the origin of any of the parameters $\boldsymbol{\sigma}^2$.

The following inequality holds

$$\sum_{k=1}^{T} \delta_{ix_k} \left(y_k - m_i\right)^2 \geq 0$$

and, as a consequence, we have

$$\exp\left\{-\frac{\sum_{k=1}^{T} \delta_{ix_k} \left(y_k - m_i\right)^2 + 2\alpha}{2\sigma_i^2}\right\} \leq \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\}$$

and

$$f\left(\boldsymbol{y}, \boldsymbol{\sigma}^2 ; \boldsymbol{m}, \boldsymbol{\theta}_X\right) \leq \sum_{\boldsymbol{x}} P\left(\boldsymbol{x} ; \boldsymbol{\theta}_X\right) \frac{1}{\sqrt{2\pi}^T} \prod_{i=1}^{N} \frac{\alpha^{\beta-1}}{\Gamma\left(\beta-1\right)} \frac{1}{\sigma_i^{\eta_i(\boldsymbol{x})+2\beta}} \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} \quad (3.10)$$

Let us now consider

$$\widetilde{\beta}\left(\boldsymbol{x}, i\right) = \beta + \frac{\eta_i\left(\boldsymbol{x}\right)}{2}$$

as a new *beta parameter* for the inverted gamma distribution, which satisfies the constraint (3.6) for a proper distribution by mean of

$$\beta > 1 \text{ and } \eta_i\left(\boldsymbol{x}\right) \geq 0 \Rightarrow \widetilde{\beta}\left(\boldsymbol{x}, i\right) > 1$$

Then, the second term of (3.10) may be written as

$$\sum_{\boldsymbol{x}} P\left(\boldsymbol{x} ; \boldsymbol{\theta}_X\right) \frac{1}{\sqrt{2\pi}^T} \prod_{i=1}^{N} \frac{\Gamma\left(\widetilde{\beta}\left(\boldsymbol{x}, i\right) - 1\right)}{\Gamma\left(\beta-1\right)} \frac{1}{\alpha^{\eta_i(\boldsymbol{x})/2}} \frac{\alpha^{\widetilde{\beta}(\boldsymbol{x},i)-1}}{\Gamma\left(\widetilde{\beta}\left(\boldsymbol{x}, i\right) - 1\right)} \frac{1}{\sigma_i^{2\widetilde{\beta}(\boldsymbol{x},i)}} \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\}$$

If we consider Property 3.2 on the inverted gamma function with the new beta parameter, *i.e.,*

$$\widetilde{f}\left(\sigma_i^2\right) = \frac{\alpha^{\widetilde{\beta}(\boldsymbol{x},i)-1}}{\Gamma\left(\widetilde{\beta}\left(\boldsymbol{x}, i\right) - 1\right)} \frac{1}{\sigma_i^{2\widetilde{\beta}(\boldsymbol{x},i)}} \exp\left\{-\frac{\alpha}{\sigma_i^2}\right\} \leq \frac{\widetilde{\beta}\left(\boldsymbol{x}, i\right)^{\widetilde{\beta}(\boldsymbol{x},i)}}{\Gamma\left(\widetilde{\beta}\left(\boldsymbol{x}, i\right) - 1\right) \alpha} \exp\left\{-\widetilde{\beta}\left(\boldsymbol{x}, i\right)\right\}$$

the likelihood function can be subsequently bounded above, obtaining

$$f\left(\boldsymbol{y},\boldsymbol{\sigma}^2;\boldsymbol{m}\right) \leq \sum_{\boldsymbol{x}} P\left(\boldsymbol{x};\boldsymbol{\theta}_X\right) \frac{1}{\sqrt{2\pi}^T}$$
$$\prod_{k=1}^{N} \frac{\Gamma\left(\widetilde{\beta}\left(\boldsymbol{x},i\right)-1\right)}{\Gamma\left(\beta-1\right)} \frac{1}{\alpha^{\eta_i(\boldsymbol{x})/2}} \frac{\widetilde{\beta}\left(\boldsymbol{x},i\right)^{\widetilde{\beta}(\boldsymbol{x},i)}}{\Gamma\left(\widetilde{\beta}\left(\boldsymbol{x},i\right)-1\right)\alpha} \exp\left\{-\widetilde{\beta}\left(\boldsymbol{x},i\right)\right\}$$
$$= \frac{1}{\sqrt{2\pi}^T} \frac{1}{\Gamma\left(\beta-1\right)^N} \sum_{\boldsymbol{x}} P\left(\boldsymbol{x};\boldsymbol{\theta}_X\right) \prod_{k=1}^{N} \frac{\widetilde{\beta}\left(\boldsymbol{x},i\right)^{\widetilde{\beta}(\boldsymbol{x},i)}}{\alpha^{1+\eta_i(\boldsymbol{x})/2}} \exp\left\{-\widetilde{\beta}\left(\boldsymbol{x},i\right)\right\} \quad (3.11)$$

By applying the following inequalities

$$\widetilde{\beta}\left(\boldsymbol{x},i\right)^{\widetilde{\beta}(\boldsymbol{x},i)} \leq \left(\beta+\frac{T}{2}\right)^{\beta+\frac{T}{2}} = \text{constant} < \infty$$
$$\alpha^{1+\eta_i(\boldsymbol{x})/2} \geq \begin{cases} \alpha & = \text{constant} < \infty, \text{ if } \alpha \geq 1 \\ \alpha^{1+T/2} & = \text{constant} < \infty, \text{ if } \alpha < 1 \end{cases}$$
$$\exp\left\{-\widetilde{\beta}\left(\boldsymbol{x},i\right)\right\} \leq 1$$

the likelihood function is finally bounded above by

$$f\left(\boldsymbol{y},\boldsymbol{\sigma}^2;\boldsymbol{m}\right) \leq \sum_{\boldsymbol{x}} P\left(\boldsymbol{x};\boldsymbol{\theta}_X\right) \text{ constant } = \text{constant} < \infty$$

The origin of any of the parameters $\boldsymbol{\sigma}^2$ no longer constitutes a point of degeneracy for the penalized likelihood. Moreover, we have found an upper bounding value for the likelihood function. □

Property 3.4 *No* $\sigma_i^2 = 0, i \in \{1 \ldots N\}$ *maximizes the penalized likelihood function.*

Proof 3.4 It is not easy to prove whether the penalized likelihood function systematically tends toward zero, as $\sigma_i^2 \to 0$. Therefore, we analyze, in the right contour of the origin of σ_i^2, the first derivative of the function with respect to σ_i^2 (more precisely the first derivative of the logarithm of the function).

By means of Property 3.6 (which is proved later on), the first derivative of the penalized log likelihood is given by

$$\frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{\sigma}^2;\boldsymbol{m}\right)}{\partial \sigma_i^2} = \frac{\alpha-\sigma_i^2\beta}{\sigma_i^4} + \sum_{k=1}^{T} P\left(X_k = x_k \mid \boldsymbol{y},\boldsymbol{\theta}^0\right) \frac{\left(y_k-m_i\right)^2-\sigma_i^2}{2\sigma_i^4} \quad (3.12)$$

In the right contour of the origin of σ_i^2 we have

$$\lim_{\sigma_i^2 \to 0^+} \frac{\alpha - \sigma_i^2 \beta}{\sigma_i^4} + \sum_{k=1}^{T} P\left(X_k = x_k \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right) \frac{\left(y_k - m_i\right)^2 - \sigma_i^2}{2\sigma_i^4} = +\infty$$

The first derivative of the likelihood function is then ensured to be positive in a contour of the origin: $\sigma_i^2 = 0, i \in \{1, \ldots, N\}$ does not maximize the penalized likelihood function (and therefore cannot be a maximum likelihood estimate). □

3.3 Likelihood Computation via the Forward - Backward Algorithm

The evaluation of the likelihood function (3.1) plays an indirect role in the computation of the maximum of the function. As discussed in [Rabiner and Juang, 1986], and proved in Appendix A, this evaluation cannot be done in a straightforward manner from the expression of the likelihood (3.1), since it would be computationally infeasible.

In order to evaluate the likelihood function, by subsequently applying the Bayes rule we obtain the following decomposition

$$f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = f\left(y_1\,;\,\boldsymbol{\theta}\right) \prod_{k=2}^{T} f\left(y_k \mid \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right) \tag{3.13}$$

In this way, we bypass the direct computation with the computation of $f\left(y_k \mid \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right)$. Such a term may be expressed as (see Appendix A for the proof)

$$f\left(y_k \mid \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right) = \sum_i \sum_j P\left(X_{k-1} = i \mid \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right) P_{ij}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_j,\sigma_j^2}\left(y_k\right) \tag{3.14}$$

The terms $\mathcal{G}_{m_j,\sigma_j^2}\left(y_k\right)$ and $P_{ij}\left(\boldsymbol{\theta}_X\right)$ can be computed in a straightforward manner (the latter once the parameterization of the Markov chain is chosen) while the probability $P\left(X_{k-1} = i \mid \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right)$ can be computed by a forward recurrence.

The forward recurrence obviously depends on the parameters $\boldsymbol{\theta} = \left\{\boldsymbol{\theta}_{Y \mid X}, \boldsymbol{\theta}_X\right\}$, therefore it depends on the parameterization of the Markov chain. We will define such a recurrence in the case of standard parameterization and in the case of telegraphic parameterization.

3.3.1 Standard Forward - Backward Algorithm

Let us consider the standard hidden Markov chain (2.12). Since the transition and initial probabilities are the characterizing parameters $\boldsymbol{\theta}_X$ of the Markov chain, we denote them explicitly with P_{ij}, p_i ; $i, j \in \{1, \ldots, N\}$.

We introduce the *Forward variable*

$$\mathcal{F}_{k-1}(i) = P\left(X_{k-1} = i \,|\, \boldsymbol{y}_1^{k-1}\,;\, \boldsymbol{\theta}\right) \tag{3.15}$$

and the *forward recurrence*

$$\begin{aligned} \mathcal{F}_1(i) &= M_1 p_i \mathcal{G}_{m_i,\sigma_i^2}(y_1) \\ \mathcal{F}_k(i) &= M_k \sum_j \mathcal{F}_{k-1}(j)\, P_{ji} \mathcal{G}_{m_i,\sigma_i^2}(y_k) \end{aligned} \tag{3.16}$$

$$\tag{3.17}$$

The M_k quantities, named the *normalizing coefficients*, are given by

$$\begin{aligned} M_1 &= \left\{ \sum_i p_i \mathcal{G}_{m_i,\sigma_i^2}(y_1) \right\}^{-1} \\ M_k &= \left\{ \sum_i \sum_j \mathcal{F}_{k-1}(j)\, P_{ji} \mathcal{G}_{m_i,\sigma_i^2}(y_k) \right\}^{-1} \end{aligned} \tag{3.18}$$

Finally, the likelihood decomposition (3.13) reads

$$f(\boldsymbol{y}\,;\,\boldsymbol{\theta}) = \prod_{k=1}^{T} M_k^{-1} \tag{3.19}$$

and its evaluation is achieved with a computational burden in the order of N^2T calculations ($3N^2$ calculations for each M_k, $3N^2$ calculation for each $\mathcal{F}_k$, with $k = 1 \ldots T$).

Similarly, we can introduce a *Backward variable*

$$\mathcal{B}_k(i) = \frac{f\left(\boldsymbol{y}_{k+1}^T \,|\, X_k = i\,;\,\boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{y}_{k+1}^T \,|\, \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}^0\right)} \tag{3.20}$$

and compute it with a backward recurrence

$$\begin{aligned}\mathcal{B}_T(i) &= 1 \\ \mathcal{B}_k(i) &= M_{k+1}\sum_j \mathcal{B}_{k+1}(j)\, P_{ij}\mathcal{G}_{m_i,\sigma_i^2}(y_k)\end{aligned} \tag{3.21}$$

The forward and backward recurrences constitute the *Forward - Backward* algorithm (in this case they constitute the standard Forward - Backward algorithm). This algorithm represents an efficient method for likelihood computation as well as for the computation of quantities appearing in likelihood maximization. A detailed proof of the recurrences and the equation (3.14), as well as a discussion on the computational burden, are in Appendix A.

3.3.2 Telegraphic Forward - Backward Algorithm

We now consider the telegraphic hidden Markov chain (2.22) with the telegraphic parameterization of the Markov chain: $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$.

By substituting such a parameterization in the forward and the backward recurrences (3.16) and (3.21) we obtain, respectively

$$\begin{aligned}\mathcal{F}_1(i) &= M_1 \frac{\mu_i/(1-\lambda_i)}{\sum_j \mu_j/(1-\lambda_j)}\mathcal{G}_{m_i,\sigma_i^2}(y_k) \\ \mathcal{F}_k(i) &= M_k\sum_j \mathcal{F}_{k-1}(j)\left[(1-\lambda_j)\mu_i + \delta_{ij}\lambda_i\right]\mathcal{G}_{m_i,\sigma_i^2}(y_k) \\ &= M_k\left\{\left[\sum_j \mathcal{F}_{k-1}(j)(1-\lambda_j)\right]\mu_i + \mathcal{F}_{k-1}(i)\,\lambda_i\right\}\mathcal{G}_{m_i,\sigma_i^2}(y_k)\end{aligned} \tag{3.22}$$

and

$$\begin{aligned}\mathcal{B}_T(i) &= 1 \\ \mathcal{B}_k(i) &= M_{k+1}\sum_j \mathcal{B}_{k+1}(j)\left[(1-\lambda_i)\mu_j + \delta_{ij}\lambda_i\right]\mathcal{G}_{m_j,\sigma_j^2}(y_{k+1}) \\ &= M_{k+1}\left\{\left[\sum_j \mathcal{B}_{k+1}(j)\,\mu_j\mathcal{G}_{m_j,\sigma_j^2}(y_{k+1})\right](1-\lambda_i) + \right. \\ &\qquad \left. + \mathcal{B}_{k+1}(i)\,\lambda_i\mathcal{G}_{m_j,\sigma_j^2}(y_{k+1})\right\}\end{aligned} \tag{3.23}$$

with the normalizing coefficients defined as

$$M_1 = \left\{ \sum_i p_i \mathcal{G}_{m_i,\sigma_i^2}(y_1) \right\}^{-1}$$

$$M_k = \left\{ \left[\sum_j \mathcal{F}_{k-1}(j)(1-\lambda_j) \right] \left[\sum_i \mu_i \mathcal{G}_{m_i,\sigma_i^2}(y_k) \right] + \sum_i \mathcal{F}_{k-1}(i) \lambda_i \mathcal{G}_{m_i,\sigma_i^2}(y_k) \right\}^{-1} \quad (3.24)$$

With the telegraphic parameterization, the transition probability between two successive different states is a separable function with respect to such states. This can be clearly seen from its expression: $P_{ij} = (1-\lambda_i)\mu_j$. This property allows us to reduce the computational burden of the recurrence: each summation over the values of a state (say at time $k-1$) is independent from the value of the successive state (say at time k). The summation in the recurrence equations (3.16) and (3.21) is then computed once for every value of the index k (being independent from the index i), and the double summation in the expression of the normalizing coefficients (3.24) is unnested. The computational burden order is then reduced from N^2T to NT calculations ($4N$ calculations for each M_k, $4N$ calculation for each $\mathcal{F}_k$ with $k = 1 \ldots T$).

3.4 Likelihood Maximization via the EM Algorithm

As discussed in the introduction of the chapter, there is no known way to find a solution to the maximization of the likelihood function in closed form. To overcome such a problem we focus on an iterative procedure known as the *Expectation Maximization* (EM) algorithm.

3.4.1 EM Algorithm

The EM algorithm is a rather general scheme for locally maximizing a likelihood function. It belongs to the class of fixed point numerical methods [Frontini, 1996]

$$\boldsymbol{\theta}^{k+1} = g\left(\boldsymbol{\theta}^k\right) \quad (3.25)$$

where $\boldsymbol{\theta}$ is the set of parameters to estimate, g is a particular function, named the *re-estimation transformation*, and k is the iteration number. After the initialization in an arbitrary point $\boldsymbol{\theta}^0 \in \Theta$, the new estimate is computed by means of (3.25), until the fixed point is attained *i.e.*, $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k$.

The theoretical foundation and properties of the EM algorithm can be found in the work [Baum *et al.*, 1970], where the EM algorithm was first developed for the statistical analysis of hidden Markov chains, and in [Dempster *et al.*, 1977] where a more generic frame is considered.

Given a set of observations $\boldsymbol{y}$, a set of parameters to estimate $\boldsymbol{\theta}$ and a corresponding likelihood function $f(\boldsymbol{y}, \boldsymbol{\theta})$, let $\boldsymbol{\theta}^k$ denote the current estimate of the parameter, and define the function $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k; \boldsymbol{y}\right)$ as

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k; \boldsymbol{y}\right) &= \int f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^k\right) \ln f\left(\boldsymbol{y}, \boldsymbol{x}\,;\, \boldsymbol{\theta}\right) d\boldsymbol{x} \\ &= E\left[\ln f\left(\boldsymbol{y}, \boldsymbol{X}\,;\, \boldsymbol{\theta}\right) \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^k\right] \end{aligned} \tag{3.26}$$

where $\boldsymbol{x}$ can be seen as an auxiliary variable which, roughly speaking, makes the "extended" likelihood $f(\boldsymbol{y}, \boldsymbol{x}\,;\, \boldsymbol{\theta})$ easier to compute.

Thus, one iteration of the algorithm is usually defined by the following two steps:

$$\begin{aligned} &\textit{Expectation (E)} && \textit{compute } Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k; \boldsymbol{y}\right) \textit{ function of } \boldsymbol{\theta} \\ &\textit{Maximization (M)} && \boldsymbol{\theta}^{k+1} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^k; \boldsymbol{y}\right) \end{aligned} \tag{3.27}$$

These two steps represent the re-estimation transformation. From now on, to preserve the generality of the iteration number, we refer to the current estimate by mean of $\boldsymbol{\theta}^0$ (as a matter of fact, note that each $\boldsymbol{\theta}^k$ can be expressed as a function of the initialization point $\boldsymbol{\theta}^0$).

The local maximization of the likelihood function is assured by the following properties of the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function of EM algorithm [Liporace, 1982]:

Property 3.5 *Any value of* $\boldsymbol{\theta}$ *such that* $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) > Q\left(\boldsymbol{\theta}^0, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ *increases the likelihood, i.e.,* $f(\boldsymbol{y}, \boldsymbol{\theta}) > f\left(\boldsymbol{y}, \boldsymbol{\theta}^0\right)$

Proof 3.5 Definition (3.26) and a logarithmic form of the Bayes rule yield:

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) - Q\left(\boldsymbol{\theta}^0, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) &= \int f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln \frac{f\left(\boldsymbol{y}, \boldsymbol{x}\,;\, \boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}, \boldsymbol{x}\,;\, \boldsymbol{\theta}^0\right)} d\boldsymbol{x} \\ &= -\int f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln \frac{f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}\right)} d\boldsymbol{x} - \ln \frac{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}\right)} \end{aligned} \tag{3.28}$$

hence

$$\ln f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}\right) - \ln f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) = Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) - Q\left(\boldsymbol{\theta}^0, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) + D\left(\boldsymbol{\theta} \| \boldsymbol{\theta}^0\right) \tag{3.29}$$

where

$$D\left(\boldsymbol{\theta}\|\boldsymbol{\theta}^0\right) = \int f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right) \ln \frac{f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\,\boldsymbol{\theta}\right)}\, d\boldsymbol{x} \tag{3.30}$$

is the Kullback-Leibler pseudo distance which is known to be non negative, as proved in [Dacunha-Castelle and Duflo, 1982]. Therefore

$$\ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) - \ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right) > Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) - Q\left(\boldsymbol{\theta}^0, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$$

and property follows. □

Property 3.6 *A parameter $\boldsymbol{\theta}$ is a critical point of the likelihood $f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right)$ if and only if it is a fixed point of the re-estimation transformation.*

Proof 3.6 Let us consider the log-likelihood function. A critical point of $\ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right)$ is characterized by $\dfrac{\partial \ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = 0$. The gradient of the log-likelihood function (under certain conditions of regularity) may be written as

$$\begin{aligned}
\frac{\partial \ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} &= \frac{1}{f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right)} \int \frac{\partial f\left(\boldsymbol{y}, \boldsymbol{x}\,;\,\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} d\boldsymbol{x} \\
&= \int f\left(\boldsymbol{y}, \boldsymbol{x}\,;\,\boldsymbol{\theta}\right) \frac{\partial \ln f\left(\boldsymbol{x} \mid \boldsymbol{y}\,;\,\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} d\boldsymbol{x} \\
&= \left.\frac{\partial Q\left(\boldsymbol{\theta}', \boldsymbol{\theta}\,;\,\boldsymbol{y}\right)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}
\end{aligned} \tag{3.31}$$

According to our standard notation we finally obtain the equivalence

$$\frac{\partial \ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)}{\partial \boldsymbol{\theta}^0} = \left.\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \tag{3.32}$$

Note that this equivalence is quite important, since it allows for the computation of the likelihood gradient via the computation of the gradient of the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ function (which, despite the gradient of the likelihood, has usually a simple expression).

From (3.31) it follows that

$$\frac{\partial \ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)}{\partial \boldsymbol{\theta}^0} = 0 \;\text{ if and only if }\; \left.\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = 0$$

The last equality is true if and only if $\boldsymbol{\theta}^0$ is a fixed point of the re-estimation transformation. □

3.4.2 EM Algorithm for Hidden Markov Chains

Let us consider the HMC model introduced in Chapter 2, where the parameters $\boldsymbol{\theta}$ to be estimated are (2.13)

$$\begin{aligned} &\textit{the parameters of the Markov chain } \boldsymbol{\theta}_X \\ &\textit{the parameters of the observable characteristic } \boldsymbol{\theta}_{Y\,|\,X} = \left\{\boldsymbol{\sigma}^2, \boldsymbol{m}\right\} \end{aligned} \tag{3.33}$$

The auxiliary variable of the EM algorithm is naturally the hidden process *i.e.,* the Markov chain. Hence, the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function is given by

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) &= E\left[\left(\ln f\left(\boldsymbol{y} \,|\, \boldsymbol{X}\,;\, \boldsymbol{\theta}_{Y\,|\,X}\right) + \ln P\left(\boldsymbol{X}\,;\, \boldsymbol{\theta}_X\right)\right) \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right] \\ &= \sum_{\boldsymbol{x}} P\left(\boldsymbol{x} \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \left\{\ln f\left(\boldsymbol{y} \,|\, \boldsymbol{x}\,;\, \boldsymbol{\theta}_{Y\,|\,X}\right) + \ln P\left(\boldsymbol{x}\,;\, \boldsymbol{\theta}_X\right)\right\} \end{aligned} \tag{3.34}$$

From this last equation (3.34), we can consider the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ function as the sum of two terms

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) &= E\left[\ln f\left(\boldsymbol{y} \,|\, \boldsymbol{X}\,;\, \boldsymbol{\theta}_{Y\,|\,X}\right) \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right] + E\left[\ln P\left(\boldsymbol{X}\,;\, \boldsymbol{\theta}_X\right) \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right] \\ &= Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) + Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) \end{aligned} \tag{3.35}$$

These terms are given by

$$\begin{aligned} Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) &= \sum_k \sum_j P\left(X_k = j \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln f\left(y_k \,|\, X_k = j\,;\, \boldsymbol{\theta}_{Y\,|\,X}\right) \\ &= \sum_{k=1}^{T} \sum_j P\left(X_k = j \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln \mathcal{G}_{m_j, \sigma_j^2}\left(y_k\right) \end{aligned} \tag{3.36}$$

and

$$\begin{aligned} Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) &= \sum_j P\left(X_1 = j \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln p_j\left(\boldsymbol{\theta}_X\right) \\ &\quad + \sum_{k=2}^{T} \sum_{i,j} P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln P_{ij}\left(\boldsymbol{\theta}_X\right) \end{aligned} \tag{3.37}$$

Once the parameterization for the Markov chain is defined, all the quantities appearing in the above terms are known, except for the probabilities $P\left(X_k = i \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right)$ and $P\left(X_{k-1} = i, X_k = j \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right)$. These probabilities can be optimally computed with the *Forward - Backward* algorithm introduced in Section 3.3. Indeed we have the relations

$$P\left(X_k = i \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \mathcal{F}_k^0(i)\, \mathcal{B}_k^0(i) \tag{3.38}$$

and

$$P\left(X_{k-1} = i, X_k = j \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \mathcal{F}_{k-1}^0(i)\, P_{ij}\left(\boldsymbol{\theta}_X^0\right) \mathcal{G}_{m_j,\sigma_j^2}(y_k)\, \mathcal{B}_k^0(i) \tag{3.39}$$

Proof of these relations is in Appendix A.

The two terms $Q_{Y \mid X}\left(\boldsymbol{\theta}_{Y \mid X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ (3.36) and $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ (3.37) depend, respectively, only on the $\boldsymbol{\theta}_{Y \mid X}$ and the parameters $\boldsymbol{\theta}_X$. Within the same iteration, the re-estimation transformations for $\boldsymbol{\theta}_{Y \mid X}$ and $\boldsymbol{\theta}_X$ are independent and obtained by separate maximizations of the two terms. Furthermore, the existence and uniqueness of the re-estimation transformation can be discussed separately.

The $Q_{Y \mid X}\left(\boldsymbol{\theta}_{Y \mid X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function fulfills such a requirement by means of the following property

Property 3.7 *The maximum of the $Q_{Y \mid X}\left(\boldsymbol{\theta}_{Y \mid X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function over the $\boldsymbol{\theta}_{Y \mid X}$ parameter space exists and is unique.*

Proof 3.7 The $Q_{Y \mid X}\left(\boldsymbol{\theta}_{Y \mid X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function is concave and tends toward $-\infty$ on the bounds of the $\boldsymbol{\theta}_{Y \mid X}$ parameter space [Liporace, 1982]. Existence and uniqueness of the maximum value follow. □

In order to discuss the existence and uniqueness of the re-estimation transformations for the parameters $\boldsymbol{\theta}_X$, we study the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function, according to the standard and the telegraphic parameterization. Then, we separately approach the computation of the re-estimation transformation for the parameters $\boldsymbol{\theta}_{Y \mid X}$ and $\boldsymbol{\theta}_X$.

3.4.2.1 $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ Function for Standard HMC

If we consider the standard parameterization, therefore we adopt a standard hidden Markov chain, from (3.37-3.39) the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ reads

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = \sum_j \mathcal{F}_1^0(j)\,\mathcal{B}_1^0(j)\ln p_j + \sum_{k=2}^{T}\sum_{i,j} \mathcal{F}_{k-1}^0(i)\,P_{ij}^0\,\mathcal{G}_{m_j^0,\sigma_j^{2^0}}(y_k)\,\mathcal{B}_k^0(i)\ln P_{ij} \quad (3.40)$$

Moreover, the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function achieves a certain properties of regularity, and as a consequence its maximum exists and is unique: the re-estimation transformation for the parameters $\boldsymbol{\theta}_X$ exist and are unique. Indeed, we prove the following property:

Property 3.8 *The maximum of the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function over the θ_X parameter space exists and is unique.*

Proof 3.8 The $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function is concave and tends toward $-\infty$ on the bounds of the parameter space [Liporace, 1982]. Existence and uniqueness of the maximum value follow. □

3.4.2.2 $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ Function for Telegraphic HMC

We consider the telegraphic parameterization $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$, and consequently the telegraphic hidden Markov model. From (3.37), the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function with such a parameterization reads

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = \sum_j P\left(X_1 = j \mid \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)\ln\frac{\mu_j/(1-\lambda_j)}{\sum_i \mu_i/(1-\lambda_i)} + \sum_{k=2}^{T}\sum_{i,j} P\left(X_{k-1}=i, X_k=j \mid \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)\ln\left((1-\lambda_i)\,\mu_j + \delta_{ij}\lambda_j\right) \quad (3.41)$$

As discussed in Section 3.3.2, about the telegraphic form of the *Forward - Backward* algorithm, the transition probability between two successive different states is a

separable function with respect to such states. By exploiting this property we obtain

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = \sum_j P\left(X_1 = j \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln \frac{\mu_j}{1-\lambda_j} - \ln \sum_i \frac{\mu_i}{1-\lambda_i}$$
$$+ \sum_{k=2}^{T} \sum_i P\left(X_{k-1} = i \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)(1-\lambda_i) + \sum_{k=2}^{T} \sum_j P\left(X_k = j \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln \mu_j$$
$$+ \sum_{k=2}^{T} \sum_j P\left(X_{k-1} = j, X_k = j \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \ln\left(\left(1-\lambda_j\right)\mu_j + \lambda_j\right) \quad (3.42)$$

where $P\left(X_k = i \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)$, $P\left(X_{k-1} = i \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)$ and $P\left(X_{k-1} = j, X_k = j \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)$ are substituted by the *Forward-Backward* equivalences (3.38) and (3.39).

In order to ease the notation, according to [Idier and Goussard, 1995] we define

$$\alpha_i^0 = \sum_{k=1}^{T} P\left(X_k = i \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \quad (3.43)$$

$$\beta_i^0 = \sum_{k=3}^{T} P\left(X_{k-1} = i \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \quad (3.44)$$

$$s_i^0 = \sum_{k=2}^{T} \sum_j P\left(X_{k-1} = j, X_k = j \mid \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right) \quad (3.45)$$

and finally the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function reads

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = \sum_j \alpha_j^0 \ln \mu_j + \sum_j \beta_j^0 \left(1-\lambda_j\right)$$
$$+ \sum_j s_j^0 \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} - \ln \sum_i \mu_i / \left(1-\lambda_i\right) \quad (3.46)$$

We can consider the telegraphic parameterization as a transformation from the standard to the telegraphic parameter space: $\{\boldsymbol{p}, \boldsymbol{P}\} \longrightarrow \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$. Such a transformation is non linear (as may be seen from (2.15) and (2.14)) and therefore Property 3.8, which guarantees the existence and uniqueness of the maximum of the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function with standard parameterization, cannot be extended to the telegraphic parameterization case. However, the following property of maximum existence holds

Property 3.9 *The maximum of the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function over the $\boldsymbol{\theta}_X$ parameter space exists.*

Proof 3.9 The $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function is bounded above and by direct calculation it is proved to tend toward $-\infty$ on the bounds of the $\boldsymbol{\theta}_X$ parameter space (such a calculation is done for the $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function introduced in Section 4.2 and the results are extended to the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function by mean of Property 4.1). Existence of the maximum value follows. □

3.5 Re-estimation Transformation for $\boldsymbol{\theta}_{Y/X}$

In order to define the re-estimation transformation for the parameters $\boldsymbol{\theta}_{Y\,|\,X}$ we maximize the $Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function via partial derivation, obtaining

$$m_i = \frac{\sum_{k=1}^{T} P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) y_k}{\sum_{k=1}^{T} P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)}$$

$$\sigma_i^2 = \frac{\sum_{k=1}^{T} P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) \left(y_k - m_i\right)^2}{\sum_{k=1}^{T} P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)}$$

which from (3.38) reads

$$m_i = \frac{\sum_{k=1}^{T} \mathcal{F}_k^0\left(i\right)^0 \mathcal{B}_k^0\left(i\right) y_k}{\sum_{k=1}^{T} \mathcal{F}_k^0\left(i\right) \mathcal{B}_k^0\left(i\right)} \tag{3.47}$$

$$\sigma_i^2 = \frac{\sum_{k=1}^{T} \mathcal{F}_k^0\left(i\right) \mathcal{B}_k^0\left(i\right) \left(y_k - m_i\right)^2}{\sum_{k=1}^{T} \mathcal{F}_k^0\left(i\right) \mathcal{B}_k^0\left(i\right)} \tag{3.48}$$

These re-estimation transformations constitute the first set of the *Baum-Welch* re-estimation formulas [Levinson *et al.*, 1983].

As discussed in Section 3.2, the likelihood function degenerates to infinity at the

origin with respect to the $\boldsymbol{\sigma}^2$ parameters. Moreover, the EM locally increases the likelihood function (Property 3.5) and its fixed point coincides with a local maximum of the likelihood (Property 3.6). Therefore, the singular points of the likelihood are an attracting domain for the EM algorithm. A penalized version of the EM algorithm is adopted to avoid the degeneracy problem.

3.5.1 Penalized EM Algorithm

Let us consider the term $Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ (3.36) of the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function (3.26), which is the only term affected by the $\boldsymbol{\theta}_{Y\,|\,X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ parameters, therefore the only term affected by the degeneracy problem.

In order to overcome the problem of the likelihood function degeneracy, discussed in Section 3.2, we have proposed a Bayesian solution (Section 3.2.1) and then we have adopted a penalized version of the likelihood function (3.7).

It is a known issue that the EM algorithm can be extended to a penalized version that converges to a local maximum of the penalized version of the likelihood function [Hero and Fessler, 1985].

Let us now consider the definition of the function $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ (3.26). If we take into account that the parameters $\boldsymbol{\sigma}^2$ are supposed to be random variables (see Section 3.2.1), the function

$$Q_{Y\,|\,X}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = E\left[\ln f\left(\boldsymbol{y}, \boldsymbol{X}\,;\,\boldsymbol{\theta}\right)\,|\,\boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right]$$

in then replaced by the "penalized" function $Q_{p_{Y\,|\,X}}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$

$$\begin{aligned} Q_{p_{Y\,|\,X}}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) &= E\left[\ln f\left(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\sigma}^2\,;\,\boldsymbol{m}, \boldsymbol{\theta}_X\right)\,|\,\boldsymbol{y}, \boldsymbol{\theta}^0\right] \\ = Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) &+ \sum_{k=1}^{T}\sum_{x_k} P\left(X_k = x_k\,|\,\boldsymbol{y}, \boldsymbol{\theta}^0\right)\ln f\left(\boldsymbol{\sigma}^2\right) \end{aligned} \tag{3.49}$$

If we consider that the parameters $\boldsymbol{\sigma}^2$ are distributed with the generic *a priori* exponential distribution (3.4), then the above "penalized" function reads

$$\begin{aligned} Q_{p_{Y\,|\,X}}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) &= Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right) + \\ &\sum_{k=1}^{T}\sum_{x_k} P\left(X_k = x_k\,|\,\boldsymbol{y}, \boldsymbol{\theta}^0\right)\left(\ln K_{norm} - \beta\ln\sigma_{x_k}^2 - \frac{\alpha}{\sigma_{x_k}^{2\gamma}}\right) \end{aligned} \tag{3.50}$$

The partial derivative with respect to the $\boldsymbol{\sigma}$ parameters is given by

$$\frac{\partial Q_{p_{Y\,|\,X}}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)}{\partial \sigma_j^2} = \frac{\alpha\gamma - \sigma_i^2\beta}{\sigma_i^{2(\gamma+1)}} + \sum_{k=1}^{T} P\left(X_k = x_k \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) \frac{(y_k - m_i)^2 - \sigma_i^2}{2\sigma_i^4} \quad (3.51)$$

The re-estimation formula of σ_i^2 is given by the equation $\dfrac{\partial Q_{p_{Y\,|\,X}}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)}{\partial \sigma_i^2} = 0$, which in order to have an explicit solution must be a first degree equation with respect to σ_i^2: $\gamma = 1$ fulfills this requirement.

The explicit σ_i^2 re-estimation formula given by the penalized version of the EM algorithm is then

$$\sigma_i^2 = \frac{\alpha + \dfrac{1}{2}\displaystyle\sum_{k=1}^{T} P\left(X_k = x_k \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)(y_k - m_i)^2}{\beta + \dfrac{1}{2}\displaystyle\sum_{k=1}^{T} P\left(X_k = x_k \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)} \quad (3.52)$$

Note that, while (3.48) can take the value zero, (3.52) leads to a $\sigma_i^2 > 0$. As a matter of fact, the numerator of (3.52) is a summation of a positive quantity and a strictly positive quantity (from (3.6) we have $\alpha > 0$).

This result was expected from the equivalence between the EM algorithm estimate and the maximum likelihood estimate: in Section 3.2 we have proved Property 3.4, which guarantees that, if the penalized likelihood function is adopted, $\sigma_i^2 = 0$ cannot be a maximum likelihood estimate.

3.6 Re-estimation Transformation for $\boldsymbol{\theta}_X$

We discuss the computation of re-estimation transformation for the parameters $\boldsymbol{\theta}_X$ of the Markov chain, in the case of standard and telegraphic parameterization.

In such a computation, we must take into account that the parameters issued from the re-estimation transformation must fulfill the probability constraint on the initial and transition probabilities of the Markov chain

$$\sum_i p_i\left(\boldsymbol{\theta}_X\right) = 1 \;,\; p_i\left(\boldsymbol{\theta}_X\right) \geq 0\,; \;\; \forall i \in \{1 \ldots N\} \quad (3.53)$$

$$\sum_j P_{ij}\left(\boldsymbol{\theta}_X\right) = 1 \;,\; P_{ij}\left(\boldsymbol{\theta}_X\right) \geq 0\,; \;\; \forall i, j \in \{1 \ldots N\} \quad (3.54)$$

3.6.1 $\boldsymbol{\theta}_X$ Parameter Estimate for Standard HMC

Let us consider the function $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ with standard parameterization (3.40). By means of a constrained maximization (via partial derivation) of this function, we obtain the re-estimation transformations for the $\boldsymbol{\theta}_X = \{\boldsymbol{p}, \boldsymbol{P}\}$ set of parameters

$$p_i = P\left(X_1 = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)$$

$$P_{ij} = \frac{\sum_{k=2}^{T} P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)}{\sum_{k=2}^{T} P\left(X_{k-1} = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)}$$

which from (3.39) read

$$p_i = \mathcal{F}_1^0(i)\,\mathcal{B}_1^0(i) \tag{3.55}$$

$$P_{ij} = \frac{\sum_{k=2}^{T} \mathcal{F}_{k-1}^0(i)\, P_{ij}^0 \mathcal{G}_{m_j^0, \sigma_j^{2\,0}}(y_k)\, \mathcal{B}_k^0(i)}{\sum_{k=1}^{T-1} \mathcal{F}_k^0(i)\,\mathcal{B}_k^0(i)} \tag{3.56}$$

These re-estimation equations constitute the second set of the *Baum-Welch* re-estimation formulas [Levinson *et al.*, 1983].

3.6.2 $\boldsymbol{\theta}_X$ Parameter Estimate for Telegraphic HMC

Let us consider the telegraphic parametrization with the "single tossing" interpretation of the transition from one state to another. As discussed in Section 2.3.1, this is the interpretation that constrains the telegraphic parameters less.

To our knowledge, pioneer work on the computation of the re-estimation transformation for telegraphic parameters must be attributed to [Goussard *et al.*, 1997]. They show that a direct maximization of the telegraphic form of $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$, under the constraints (2.24) and (2.26), yields an untractable expression that fails to give an explicit re-estimation transformation. This problem is attributed to the contribution of the last term of (3.40).

The loss of the explicitness of the re-estimation formulas may jeopardize the choice of the EM algorithm. In in order to remain within the EM algorithm frame, and therefore benefit from its attractive features, in [Goussard *et al.*, 1997] the authors propose to modify the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ function. Such a modification leads to the definition of

a *telegraphic EM algorithm*. The next chapter will be devoted to such a definition.

We may believe the loss of explicitness to be stricly related to the choice of telegraphic parameters but, actually, it is an intrinsic outcome of the reversibility property (induced by such a choice), as pointed out in the following section [2].

3.6.2.1 Reversibility Constraint

According to Definition 2.2, a homogeneous Markov chain with characterizing parameters denoted with $\boldsymbol{\theta}_X$, is said to be reversible if and only if it satisfies the following condition:

$$p_i\left(\boldsymbol{\theta}_X\right) P_{ij}\left(\boldsymbol{\theta}_X\right) = p_j\left(\boldsymbol{\theta}_X\right) P_{ji}\left(\boldsymbol{\theta}_X\right) \tag{3.57}$$

The reversibility constraint is not easy to introduce in a constrained maximization procedure, despite the probability constraints (3.53) and (3.54).

To overcome such a problem we consider the equivalent reversibility condition

$$\tilde{P}_{ij}\left(\boldsymbol{\theta}_X\right) = \tilde{P}_{ji}\left(\boldsymbol{\theta}_X\right) \tag{3.58}$$

where

$$\tilde{P}_{ij}\left(\boldsymbol{\theta}_X\right) = P\left(X_k = i, X_{k-1} = j\,;\,\boldsymbol{\theta}_X\right)\ \forall k = 2 \ldots T \tag{3.59}$$

and in order to have it intrinsically satisfied, we parameterize the Markov Chain with the parameters $\tilde{\boldsymbol{\theta}}_X = \left\{\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{P}}\right\}$. The parameters $\tilde{\boldsymbol{P}} = \left\{\tilde{P}_{ij}\right\}$ are defined in (3.59) and the parameters $\tilde{\boldsymbol{p}} = \{\tilde{p}_i\}$ (the stationary probabilities) are defined as

$$\tilde{p}_i = \sum_j \tilde{P}_{ij} = \sum_j \tilde{P}_{ji} \tag{3.60}$$

In order to apply the EM algorithm we must re-write $Q_X\left(\tilde{\boldsymbol{\theta}}_X, \tilde{\boldsymbol{\theta}}^0; \boldsymbol{y}\right)$, which is the only term of the function $Q\left(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}^0; \boldsymbol{y}\right)$ affected by the Markov chain parameters, by taking into account the new parameterization.

[2] All the computations in Section 3.6.2.1 have been developed by FrĆderic Champagnat, Groupe ProblŁmes Inverses - Laboratoire des Signaux et SystŁmes

If we express

$$\ln P\left(\boldsymbol{x}\,;\,\tilde{\boldsymbol{\theta}}_X\right) = \sum_{k=2}^{T} \ln \tilde{P}_{x_{k-1}x_k} - \sum_{k=2}^{T-1} \ln \tilde{P}_{x_k} \tag{3.61}$$

from the definition (3.37) of $Q_X\left(\tilde{\boldsymbol{\theta}}_X, \tilde{\boldsymbol{\theta}}^0; \boldsymbol{y}\right)$ we obtain

$$Q_X\left(\tilde{\boldsymbol{\theta}}_X, \tilde{\boldsymbol{\theta}}^0; \boldsymbol{y}\right) = \sum_{k=2}^{T} \sum_{i,j} P\left(X_{k-1} = i, X_k = j \mid \boldsymbol{y}\,;\,\tilde{\boldsymbol{\theta}}^0\right) \ln \tilde{P}_{ij} - \sum_{k=2}^{T-1} \sum_{j} P\left(X_1 = j \mid \boldsymbol{y}\,;\,\tilde{\boldsymbol{\theta}}^0\right) \ln \tilde{p}_j \tag{3.62}$$

With such a function, the EM estimates of the parameters $\tilde{\boldsymbol{\theta}}_X = \left\{\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{P}}\right\}$, issued by the re-estimation transformation, intrinsically satisfy the reversibility constraint.

The re-estimation transformations are the solution of the equation

$$\frac{\partial Q_X\left(\tilde{\boldsymbol{\theta}}_X, \tilde{\boldsymbol{\theta}}^0; \boldsymbol{y}\right)}{\partial \tilde{\boldsymbol{\theta}}_X} = 0$$

which does not yield any tractable expression.

Since such a choice of parameterization, (3.59) and (3.60), is rather general, the non explicitness can be considered a consequence of the way we have defined the quantities (3.60) and (3.61). Therefore, it is a consequence of the property of reversibility.

Chapter 4

Telegraphic EM Algorithm for the Estimation of $\boldsymbol{\theta}_X$

4.1 Introduction

In the previous chapter we introduced the maximum likelihood estimation of the hidden Markov chain parameters via the EM algorithm. With such an algorithm the $\boldsymbol{\theta}_{Y\,|\,X}$ parameters and the standard $\boldsymbol{\theta}_X$ parameters are re-estimated by means of the Baum-Welch formulas (Section 3.5 and Section 3.6.1), while the telegraphic $\boldsymbol{\theta}_X$ parameters fail to have explicit re-estimation equations (Section 3.6.2). In order to obtain explicit re-estimation equations for the telegraphic parameters, in [Goussard *et al.*, 1997], and previously in [Idier and Goussard, 1995], the authors propose to substitute the function $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ with a new one, denoted as $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$. In this way they define a new EM algorithm, based on the maximization of the auxiliary function

$$R\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) + R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) \tag{4.1}$$

instead of

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) + Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) \tag{4.2}$$

We refer to this new iterative procedure as a *telegraphic EM algorithm* (TEM).

In this chapter we propose a study of the new auxiliary function. Our purpose is to find properties which may be sufficient to ensure the local maximization of the likelihood function via the telegraphic EM algorithm. To our knowledge, such a study is not present in the scientific literature and constitutes an original contribution.

Then, according to [Goussard *et al.*, 1997], we expose the re-estimation transformations for the telegraphic parameter, issued from the maximization of the auxiliary function of the telegraphic EM algorithm.

4.2 $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ Function for the Telegraphic EM Algorithm

In Section 3.4.2.2 the function $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ with the telegraphic parameterization (3.46) has been defined as

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = \sum_j \alpha_j^0 \ln \mu_j + \sum_j \beta_j^0 \left(1-\lambda_j\right) + \sum_j s_j^0 \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} - \ln \sum_i \mu_i / \left(1-\lambda_i\right) \quad (4.3)$$

In [Goussard *et al.*, 1997], they consider the last term of (4.3) to be responsible for the non explicitness of the re-estimation transformation. However, they also remark that as the number of observations increases, such a term becomes small with respect to the other ones, and therefore may be neglected. Before we approximate the auxiliary function we should note that, as a consequence of the reversibility property of the telegraphic Markov chain, such a function is invariant by reversion of the time index of the chain (which in our notation is the index k). This means that if we go through the chain in a backward manner *i.e.,* we substitute the index k with the index $T-k+1$, the auxiliary function does not change.

By maintaining the invariance by reversion of the time index of the chain, the approximation of (4.3) yields the new function

$$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = \sum_j \frac{\alpha_j^0 + \beta_j^0}{2} \ln \mu_j \left(1-\lambda_j\right) + s_j^0 \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} \quad (4.4)$$

with

$$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) \cong Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) \quad (4.5)$$

Additionally to (3.43-3.45), we define

$$\gamma_i^0 = \frac{\alpha_i^0 + \beta_i^0}{2} - s_i^0 \quad (4.6)$$

The above quantity is positive *i.e.,* $\gamma_i^0 \geq 0\ \forall i \in \{1, \ldots, N\}$, as it is proved in [Idier and Goussard, 1995] by exploiting the property of reversibility of the telegraphic Markov chain.

If we adopt this new definition (4.6) and we take into account, from (4.4), that $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ is a summation of independent terms, we can write

$$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = \sum_j R_{Xj}\left(\boldsymbol{\theta}_{Xj}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$$

where

$$R_{Xj}\left(\boldsymbol{\theta}_{Xj}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = \gamma_j^0 \ln \mu_j \left(1-\lambda_j\right) + s_j^0 \ln\left(\left(1-\lambda_j\right)\mu_j + \lambda_j\right) \tag{4.7}$$

and

$$\boldsymbol{\theta}_{Xj} = \{\mu_j, \lambda_j\} \tag{4.8}$$

Now let us examine the properties of the new algorithm to check whether it converges to a local maximum of the likelihood function.

For the classic EM algorithm, introduced in Section 3.4.1, the local maximization of the likelihood function was ensured by Property 3.6 and Property 3.5. Moreover Property 3.7 guaranteed the existence and uniqueness of the re-estimation transformations for the parameter $\boldsymbol{\theta}_{Y\,|\,X}$ of the observable characteristic, and Property 3.9 guaranteed the existence of the re-estimation transformations for the parameters $\boldsymbol{\theta}_X$ of the telegraphic Markov chain.

All these properties are strictly related to the function $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ (which is considered as a function of $\boldsymbol{\theta}$). As a consequence of the fact that the latter is composed by two mutually independent terms ($Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ which depends only on $\boldsymbol{\theta}_{Y\,|\,X}$ and $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ which depends only on $\boldsymbol{\theta}_{Y\,|\,X}$), Property 3.6 and Property 3.5 apply independently on the two terms.

Note that $R\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ differs from $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ only by the term on the $\boldsymbol{\theta}_X$ parameters. Therefore, with respect to the $\boldsymbol{\theta}_{Y\,|\,X}$ parameters, Property 3.6 and Property 3.5 also apply to the function $R\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$. Thus, the analysis of the algorithm can be limited to a study of the characteristics of $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$. Such a study yields the following three properties:

Property 4.1 *The function* $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ *with telegraphic parameterization is inferior to the function* $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ *i.e.,*

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) \leq R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$$

Proof 4.1 The function $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ with telegraphic parameterization (3.46) may

be written as

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = \sum_j \gamma_j^0 \ln \mu_j \left(1-\lambda_j\right) + \sum_j \frac{\alpha_j^0 + \beta_j^0}{2} \ln \frac{\mu_j}{1-\lambda_j} + \sum_j s_j^0 \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} - \ln \sum_i \frac{\mu_i}{1-\lambda_i} \quad (4.9)$$

Let us now define

$$\frac{\overline{\mu}}{1-\overline{\lambda}} = \max_{i \in \{1 \ldots N\}} \frac{\mu_i}{1-\lambda_i}$$

From the inequalities

$$\sum_j \frac{\alpha_j^0 + \beta_j^0}{2} \ln \frac{\mu_j}{1-\lambda_j} \leq \ln \frac{\overline{\mu}}{1-\overline{\lambda}} \sum_j \frac{\alpha_j^0 + \beta_j^0}{2} \ln \frac{\overline{\mu}}{1-\overline{\lambda}}$$

and

$$-\ln \sum_i \frac{\mu_i}{1-\lambda_i} \leq -\ln \frac{\overline{\mu}}{1-\overline{\lambda}}$$

we obtain

$$\sum_j \frac{\alpha_j^0 + \beta_j^0}{2} \ln \frac{\mu_j}{1-\lambda_j} - \ln \sum_i \frac{\mu_i}{1-\lambda_i} \leq 0$$

Therefore, we have

$$Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) \leq \sum_j \gamma_j^0 \ln \mu_j \left(1-\lambda_j\right) + \sum_j s_j^0 \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} = R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$$

□

Property 4.2 *$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ is a concave function on the $\boldsymbol{\theta}_X$ variable.*

Proof 4.2 The function $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ is defined as (4.7)

$$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = \sum_j R_{X\,j}\left(\boldsymbol{\theta}_{X\,j}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$$

and the Hessian of each $R_{Xj}\left(\boldsymbol{\theta}_{Xj}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ is given by

$$\boldsymbol{H}_{R_{Xj}}\left(\boldsymbol{\theta}_{Xj}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = \begin{bmatrix} -\frac{\gamma_i^0}{\mu_j^2} - \frac{s_j^0(1-\lambda_j)^2}{(\mu_j(1-\lambda_j)+\lambda_j)^2} & -\frac{1}{(\mu_j(1-\lambda_j)+\lambda_j)^2} \\ -\frac{1}{(\mu_j(1-\lambda_j)+\lambda_j)^2} & -\frac{\gamma_j^0}{(1-\lambda_j)^2} - \frac{s_j^0(1-\mu_j)^2}{(\mu_j(1-\lambda_j)+\lambda_j)^2} \end{bmatrix}$$

Since from definition (3.45) and (4.6) we have $s_i \geq 0$ and $\gamma_i^0 \geq 0\ \forall i \in \{1, \ldots, N\}$, all the minors of the matrix are negative. Moreover, from the definition of the domains of λ_i and μ_j, the product of the terms on the main diagonal of the matrix can be shown to be inferior to the product of the terms in the anti diagonal: the determinant is then ensured to be negative.

These are sufficient conditions to prove that the Hessian matrix is defined negative. $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ is then a concave function, and since the concavity property is stable through addition, the concavity of the whole $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ function follows. □

Property 4.3 *$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$, as a function of $\boldsymbol{\theta}_X$, tends toward $-\infty$ on the bounds of the parameters space.*

Proof 4.3 The domain of the $\boldsymbol{\theta}_X$ parameters is defined by the following relations

$$0 < \mu_i < 1\,;\ \forall i \in \{1, \ldots, N\}\,,\ \text{with}\ \sum_{j=1}^{N} \mu_j = 1 \tag{4.10}$$

$$-\frac{\mu_i}{1-\mu_i} < \lambda_i < 1\,;\ \forall i \in \{1, \ldots, N\} \tag{4.11}$$

In figure 4.3 we have represented the domain of the parameters $\boldsymbol{\mu} = \{\mu_i\,;\, i = 1 \ldots N\}$ in the simple case of a three parameters set *i.e.*, $\boldsymbol{\mu} = \{\mu_1, \mu_2, \mu_3\}$, and the domain of each parameter λ_i as a function of the parameter μ_i.

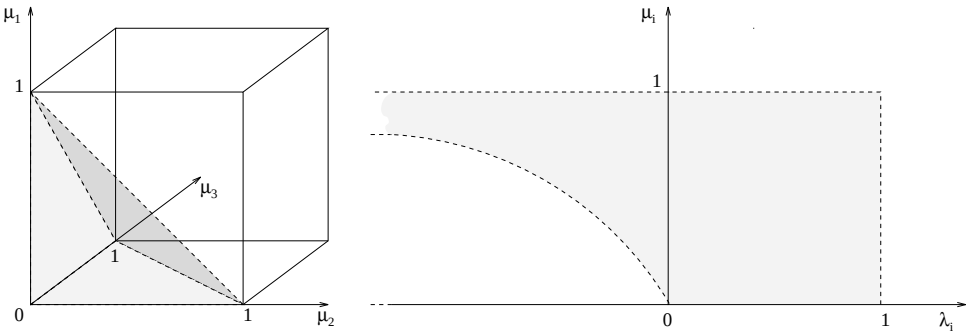

Figure 4.1: Domains of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$

The bounds of the parameters space are attained in the following cases:

1) $\lambda_i \to 1$, $i \in \{1, \ldots, N\}$
we have

$$\lim_{\lambda_i \to 1} \gamma_i^0 \ln \mu_i (1 - \lambda_i) = -\infty$$
$$\lim_{\lambda_i \to 1} s_i^0 \ln (\mu_i (1 - \lambda_i) + \lambda_i) = 0$$

therefore,

$$\lim_{\lambda_i \to 1} R_X \left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = -\infty; \quad \forall \mu_i \in \overline{\Theta}, \quad \forall i \in \{1, \ldots, N\}$$

2) $\lambda_i \to -\mu_i / (1 - \mu_i)$, $i \in \{1, \ldots, N\}$
we have

$$\lim_{\lambda_i \to -\frac{\mu_i}{1-\mu_i}} \gamma_i^0 \ln \mu_i (1 - \lambda_i) = -\gamma_i^0 \ln (1 - \mu_i) + \gamma_i^0 \ln \mu_i \tag{4.12}$$
$$\lim_{\lambda_i \to -\frac{\mu_i}{1-\mu_i}} s_i^0 \ln (\mu_i (1 - \lambda_i) + \lambda_i) = -\infty \tag{4.13}$$

We must guarantee that the term (4.12), or the sum of such terms, stays strictly inferior to $+\infty$. Intuitively this condition may be unsatisfied for $\mu_i \to 1$ (note that from definition (4.6), γ_i^0 is a bounded constant and it does not influence the limit). Therefore, this case needs a precise analysis.

Let us consider the constraint over the $\boldsymbol{\mu}$ parameters: $\sum_{j=1}^{N} \mu_j = 1$. This relation implies that if $\mu_i \to 1 \Rightarrow \mu_j \to 0$, $\forall j \neq i$. In a contour of $\mu_i = 1$, we may assume that $\forall j \neq i$, $\mu_j \to 0$ along the generic line given by the equation

$$\mu_j = a_j (1 - \mu_i) \tag{4.14}$$

where $a_j > 0$. In order to provide a graphical representation of the above situation, we consider the simple case of a three parameter set $\boldsymbol{\mu} = \{\mu_1, \mu_2, \mu_3\}$, where we assume the parameter μ_i to be the parameter μ_1. In figure 4.3 we have represented the domain of the parameters $\boldsymbol{\mu} = \{\mu_1, \mu_2, \mu_3\}$ in a contour of $\mu_1 = 1$, and the generic lines along which $\mu_j \to 0$; $j = 2, 3$.

If we substitute the line equation (4.14) into the summation of the terms (4.12), we

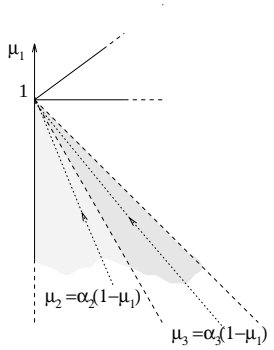

Figure 4.2: Limit of μ

obtain

$$\sum_{j=1}^{N} -\gamma_j^0 \ln\left(1-\mu_j\right) + \gamma_j^0 \ln \mu_j = \left(\sum_{j\neq i} \gamma_j^0 - \gamma_i^0\right) \ln\left(1-\mu_i\right) + \sum_{j\neq i} \gamma_j^0 \ln a_j - \sum_{j\neq i} \ln\left(1 - a_j\left(1-\mu_i\right)\right) + \ln \mu_i$$

The last three terms are not influent since one of them is a constant and the others tend to zero as $\mu_i \to 1$. In [Idier and Goussard, 1995] the authors have proved that

$$\sum_{j=1}^{N} \gamma_j^0 > 2\gamma_i^0\,;\ \ \forall i \in \{1,\ldots,N\}$$

therefore,

$$\lim_{\mu_i \to 1} \left(\sum_{j\neq i} \gamma_j^0 - \gamma_i^0\right) \ln\left(1-\mu_i\right) = -\infty$$

Finally

$$\lim_{\lambda_i \to -\frac{\mu_i}{1-\mu_i}} R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right) = -\infty\,;\ \ \forall \mu_i \in \overline{\Theta},\ \ \forall i \in \{1,\ldots,N\}$$

3) $\lambda \in \Theta,\ \forall i \in \{1,\ldots,N\}$
we consider only the bound values of the parameters $\boldsymbol{\mu}$.

3.1) If $\mu_i \to 0$ we have $\lim_{\mu_i \to 0} R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = -\infty$. Therefore

$$\lim_{\mu_i \to 0} R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = -\infty, \quad \forall i \in \{1, \ldots, N\}$$

3.2) If $\mu_i \to 1$ we obtain

$$R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = \gamma_i^0 \left(1 - \lambda_i\right) < \infty$$

As $\mu_i \to 1 \Rightarrow \mu_j \to 0 \; \forall j \neq i$, and by the results obtained in **2)**,

$$\lim_{\mu_j \to 0} R_{Xj}\left(\boldsymbol{\theta}_{Xj}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = -\infty ; \quad \forall j \neq i$$

Therefore,

$$\lim_{\mu_j \to 0} R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) = -\infty, \quad \forall i \in \{1, \ldots, N\}$$

□

Figure 4.3 shows one of the functions $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ and, on the xy plane, the domain of the parameters λ_i and μ_i. We can observe the concavity and the values of the function on the bounds of the parameter space. The value of the $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ function on the bound $\mu_i = 1$ appears finite. This is not contradictory with respect to Property 4.3. As explained above, as $\mu_i \to 1$, the functions $R_{Xj}\left(\boldsymbol{\theta}_{Xj}, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right) \; ; \; \forall j \neq i$ tend toward $-\infty$. Therefore, $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ tends toward $-\infty$ as whole summation.

Property 4.2 and Property 4.3 guarantee the existence and uniqueness of the re-estimation transformation for the parameters $\boldsymbol{\theta}_X$, within the framework of the telegraphic EM algorithm, while Property 4.1 and Property 4.3 allow to prove Property 3.9 of Section 3.4.2.2 (values on the bounds of the parameter space of the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ function with telegraphic parameterization). However, they are not sufficient to ensure that the telegraphic EM algorithm locally maximizes the likelihood function.

With a small set of observations, a counterexample has shown a situation in which the likelihood function does not increase and the algorithm stalls in a loop (see Section 7.4). We can then affirm that sufficient properties, which guarantee the convergence of the telegraphic EM algorithm to a local maximum of the likelihood function, do not exist.

Nevertheless, with a large set of observations, the algorithm seems not to suffer of such a problem. This is coherent with what observed in [Goussard *et al.*, 1997]: as the number of observations increases, the $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ function approaches the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ function (the term that has been neglected in $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ in order to obtain $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 ; \boldsymbol{y}\right)$ becomes small with respect to the other terms). We can then

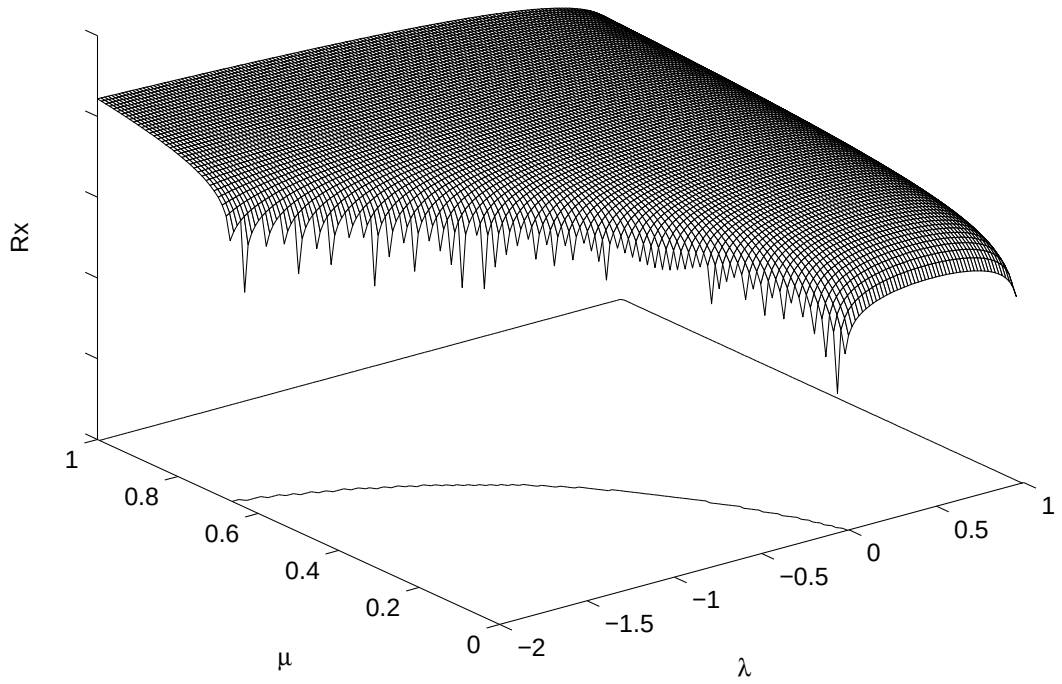

Figure 4.3: $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ Function

assume that, with a large set of observations, the properties of the $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ function may be extended to the $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}\right)$ function.

The fact of not being able to guarantee the local maximization of the likelihood function may be a delicate problem. Moreover, the EM algorithm already suffers of a slow convergence rate as discussed in [Campillo and Le Gland, 1989] and [Meilijson, 1989], and even when the local maximization is theoretically ensured, convergence time may be too high to attain the local maximum. This is clearly shown by the gradient of the likelihood: after a certain number of iterations the gradient stabilizes, but to a value different than zero (see Section 7.4).

We propose a mixed EM-Gradient technique in order to solve the problem of a possible slow convergence rate and in order to overcome the absence of properties which guarantee the convergence to a local maximum. This mixed technique is described in Appendix B, and its performances are analyzed in Section 7.4.

4.3 Re-estimation Transformation for $\boldsymbol{\theta}_X$ Parameters

The re-estimation formulas for the parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$, within the framework of the telegraphic EM algorithm, are obtained by means of the maximization of the $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ function.

Such a maximization is not easy and, to our knowledge, it has been firstly approached by [Idier and Goussard, 1995]. In the following we expose the way they have obtained the re-estimation formulas.

From (4.7), the maximization of the function $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ can be computed separately on each term $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$, via partial derivation with respect to the parameters $\boldsymbol{\theta}_{Xi} = \{\mu_i, \lambda_i\}$.

We consider the "single toss" interpretation of the telegraphic Markov chain (see Section 2.3.1), and therefore we constraint the parameters $\boldsymbol{\mu}$ to fulfill

$$\sum_{j=1}^{N} \mu_j = 1 \;,\; \mu_i \geq 0\,;\;\; \forall i \in \{1 \ldots N\} \tag{4.15}$$

and the parameters $\boldsymbol{\lambda}$ to fulfill

$$-\frac{\mu_i}{1-\mu_i} \leq \lambda_i \leq 1\,;\;\; \forall i \in \{1 \ldots N\} \tag{4.16}$$

Direct maximization of $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ with respect to the parameter λ_i yields the re-estimation transformation

$$\widehat{\lambda}_i = \frac{s_i^0 / \left(s_i^0 + \eta_i^0\right) - \mu_i}{1-\mu_i} \tag{4.17}$$

which depends on μ_i. From the definitions (3.43-3.45) and (4.6), the estimate $\widehat{\lambda}_i$ issued from (4.17) automatically fulfills the inequality constraint (4.16).

The substitution of such an estimate in $R_{Xi}\left(\boldsymbol{\theta}_{Xi}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ gives

$$R_{Xi}\left(\mu_i, \widehat{\lambda}_i, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right) = \gamma_i^0 \ln \frac{\mu_i}{1-\mu_i} + const \tag{4.18}$$

The constrained maximization of (4.18) with respect to the parameter μ_i (under the constraints (4.15)) is a more complex problem.

In [Idier and Goussard, 1995] the authors first constrain such a maximization to

$\mu_i \geq 0$. By applying a Lagrangian maximization method they obtain

$$\nu\mu_i^2 - \nu\mu_i^2 + \gamma_i^0 = 0$$

where ν is the Lagrange multiplier.

The indeterminacy on the possible solutions of the above equations is reduced by the equality constraint $\sum_{j=1}^{N} \mu_j = 1$. Finally, the equation has a unique solution $\widehat{\mu}_i\,(\widehat{\nu})$. The value of $\widehat{\nu}$ is not explicit: In [Idier and Goussard, 1995] ν is initially framed in a strict interval and then $\widehat{\nu}$ is compute by an interpolation method.

The re-estimates of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ exist and are unique: note that the uniqueness and existence was expected from the uniqueness and existence of the maximum value of the function $R_{X\,i}\left(\boldsymbol{\theta}_{X\,i}, \boldsymbol{\theta}^0\,;\, \boldsymbol{y}\right)$ (this last property is a consequence of Property 4.2 and Property 4.3 on the concavity and the bound values of the function, respectively).

Chapter 5

Bidimensional Hidden Markov Models

5.1 Introduction

In this chapter we introduce the bidimensional hidden Markov models. Their definition, as well as their meaning, may be considered as an extension of the unidimensional hidden Markov models introduced in Chapter 2: they are doubly stochastic processes, with an underlying bidimensional stochastic process $\boldsymbol{X} = \{X_{r,c}\}_{(r,c)\in\Lambda}$ that is not observable (hidden) but can only be observed through another bidimensional stochastic process $\boldsymbol{Y} = \{Y_{r,c}\}_{(r,c)\in\Lambda}$ that produces the set of observations $\boldsymbol{y} = \{y_{r,c}\}_{(r,c)\in\Lambda}$.

A bidimensional rectangular lattice $\Lambda = \{(r,c) \mid l = 1 \ldots R, c = 1 \ldots C\}$ substitutes the observation time interval $1 \ldots T$, and the points of the lattice (r,c), named *sites* and denote with a couple of spatial indexes, substitute the time instants k.

Such a lattice is supposed to be finite, with R rows and C columns, and to be oriented as is figure 5.1

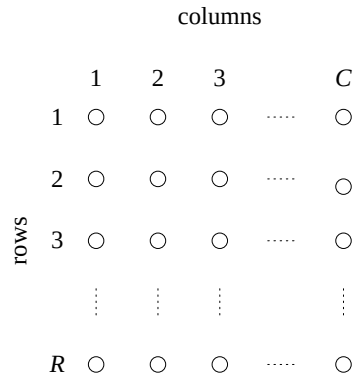

Figure 5.1: Rectangular Lattice

To understand the concept of the bidimensional hidden Markov models lets go back to the example we have used to explain the unidimensional hidden Markov models. This example applies to the multidimensional case, with the difference that behind the

curtain we have now RC coin tossing experiments and you are told the results of the RC experiments at the same time.

In our bidimensional hidden Markov Model, the sample spaces of the two processes are defined according to the assumptions and the notations of the unidimensional case: each random variable of the observable and hidden stochastic processes is sampled on a finite continuous set $\mathcal{O} \subseteq \mathbb{R}$ and on a finite (numerable) state set $\mathcal{S} = \{m_1, \ldots, m_N\}$, respectively.

The observable process $\boldsymbol{Y} = \{Y_{r,c}\}_{(r,c)\in\Lambda}$ is still considered to be related to the hidden one $\boldsymbol{X} = \{X_{r,c}\}_{(r,c)\in\Lambda}$ by independent Gaussian distributions

$$f\left(Y_{r,c} \,|\, X_{r,c} = x_{r,c}\right) = \mathcal{G}_{m_{x_{r,c}},\sigma^2_{x_{r,c}}}\left(Y_{r,c}\right) \;;\; x_{r,c} \in \{1, \ldots, N\}, \quad (r,c) \in \Lambda \tag{5.1}$$

where $\mathcal{G}_{m_{x_{r,c}},\sigma^2_{x_{r,c}}}\left(Y_{r,c}\right)$ is defined as in (2.7).

The joint distribution of the observable process given the realization of the hidden process is now

$$f\left(\boldsymbol{Y} \,|\, \boldsymbol{x}\right) = \prod_{r=1}^{L}\prod_{c=1}^{C} \mathcal{G}_{m_{x_{r,c}},\sigma^2_{x_{r,c}}}\left(Y_{r,c}\right) \tag{5.2}$$

This is the observable characteristic of the model and as in the unidimensional case it depends on $2N$ parameters denoted as $\boldsymbol{\theta}_{Y\,|\,X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$, where the parameters $\boldsymbol{m} = \{m_i \,|\, m_i \in \mathcal{S}\,;\, i = 1 \ldots N\}$ are the elements of the state space (the values of the states), and the parameters $\boldsymbol{\sigma}^2 = \left\{\sigma_i^2 \,|\, \sigma_i^2 \in \mathbb{R}_+^*\,;\, i = 1 \ldots N\right\}$ are the variances of the Gaussian distributions.

The crucial difference between the unidimensional hidden Markov models and the bidimensional hidden Markov models resides in the structure of the hidden process. In the unidimensional case the hidden process is a Markov chain (2.2), while in the bidimensional case is a *Markov random field* (MRF) [Brémaud, 1997].

Definition 5.1 Markov Random Field
A Markov random field is a bidimensional stochastic process $\boldsymbol{X} = \{X_{r,c}\}_{(r,c)\in\Lambda}$, defined over a lattice Λ (we adopt the finite rectangular lattice previosly introduced). Each random variable of the bidimensional stochastic process $X_{r,c}\,;\, (r,c) \in \Lambda$ is sampled over a common finite state space (we adopt the state space $\mathcal{S}$), and satisfies the Markov property. In the unidimensional case, such a property is defined as in (2.1), while for a bidimensional stochastic process it can be defined as

$$P\big(X_{r,c} = x_{r,c} \,|\, X_{m,n} = x_{m,n}\,;\, (m,n) \in \Lambda / \{(r,c)\}\big) = P\big(X_{r,c} = x_{r,c} \,|\, X_{m,n} = x_{m,n}\,;\, (m,n) \in \mathcal{N}(r,c)\big)\,;\quad \forall\,(r,c) \in \Lambda \tag{5.3}$$

$\Lambda / \{(r,c)\}$ is the rectangular lattice without the site (r,c), and $\mathcal{N}(r,c)$ is the *neighborhood system* of site (r,c). The neighborhood system of a site (r,c) can be defined as

$$\mathcal{N}(r,c) = \left\{(m,n) \in \Lambda \,|\, 0 < (m-r)^2 + (n-c)^2 \leq d\right\} \tag{5.4}$$

The parameter d determines the *order* of the neighborhood system, and therefore the order of the Markov property: the Markov random field is then said to be of *order* d.

In figure 5.2 we provide a graphical representation of a neighborhood system of order 1 (on the left) and a neighborhood system of order 2 (on the right). This second neighborhood system is the one adopted for our Markov random field (which is then called *second-order Markov random field*).

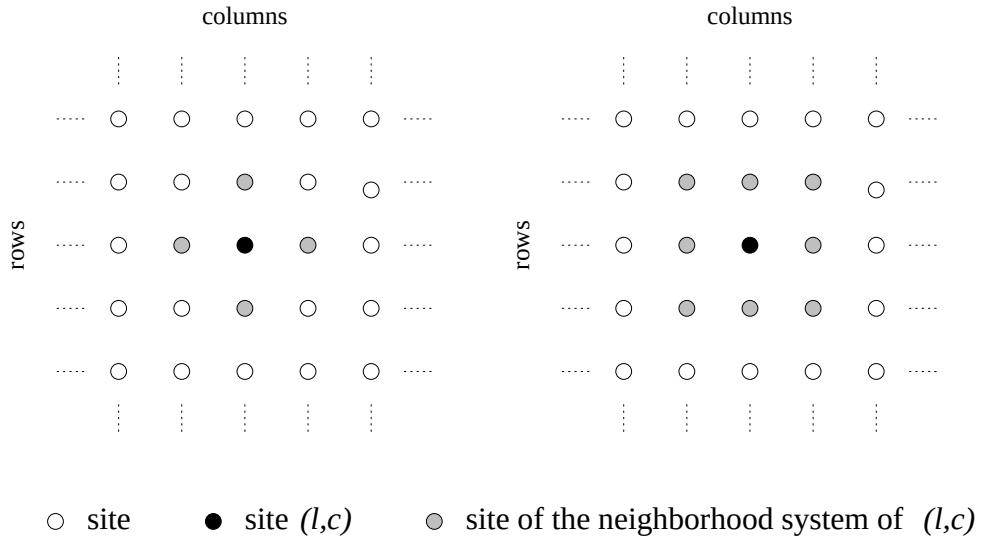

Figure 5.2: Neighborhood Systems of Order 1 and 2

The probability

$$P\left(X_{r,c} = x_{r,c} \,|\, X_{m,n} = x_{m,n}\,;\, (m,n) \in \mathcal{N}(r,c)\right) \tag{5.5}$$

is named the *local characteristic* of the Markov random field at site (r,c).

Under certain conditions, the set of the local characteristics defines the Markov random field [Brémaud, 1997]. △

The definition of the parameters that characterize a general Markov random field requires particular assumptions and the theory of Gibbs fields. We avoid this complication by specifically base our model on Pickard random fields (PRFs) [Pickard, 1980],

and by considering a special set of parameters that are sufficient to characterize the model within our applied issue framework.

Note that, despite the unidimensional case, it is difficult to provide a graphical representation of a bidimensional hidden Markov model. Thus, we appeal to the reader's immagination.

5.2 Pickard Random Fields

Definition 5.2 Pickard Random Field
A Pickard random field is a second order Markov random field $\boldsymbol{X} = \{X_{r,c}\}_{(r,c)\in\Lambda}$ with probability distribution of the field given by

$$P(\boldsymbol{x}) = \tau(x_{1,1}) \prod_{r=2}^{R} \tau(x_{r,1} \mid x_{r-1,1}) \prod_{c=2}^{C} \tau(x_{1,c} \mid x_{1,c-1}) \prod_{r=2}^{R} \prod_{c=2}^{C} \tau(x_{r,c} \mid x_{r-1,c-1}, x_{r-1,c}, x_{r,c-1}) \quad (5.6)$$

τ is a measure defined on a generic square cell of four sites *i.e.*, $\tau \begin{pmatrix} A=a & B=b \\ C=c & D=d \end{pmatrix}$, that satisfies the following conditions of symmetry and independence:

$$\tau(B \mid A, C) = \tau(B \mid A), \quad \tau(C \mid A, B) = \tau(C \mid A)$$

and

$$\begin{gathered} \tau(B \mid D, C) = \tau(B \mid D), \quad \tau(C \mid D, B) = \tau(C \mid A) \\ \text{or} \\ \tau(A \mid B, D) = \tau(A \mid B), \quad \tau(D \mid B, A) = \tau(D \mid B) \\ \tau(A \mid C, D) = \tau(A \mid C), \quad \tau(D \mid C, A) = \tau(D \mid C) \end{gathered}$$

△

As a consequence of the distinctive characteristics of the measure τ, the Pickard random fields benefit from the following properties (see [Pickard, 1980] for the proofs):

Property 5.1 *The Pickard random fields are stationary, and therefore homogeneous, Markov random field.*

Definition 5.3 Stationarity and Homogeneity Properties of a Markov Random Field
Let $\mathcal{T}$ be the set of translations on $\mathbb{Z}^2$. A Markov random field (and more generically a stochastic process) $\boldsymbol{X} = \{X_{r,c}\}_{(r,c)\in\Lambda}$ is said to be stationary on the finite rectangular lattice Λ if and only if $\forall\zeta \in \mathcal{T}$, $\forall A \subset \Lambda$ such that $\zeta(A) \subset \Lambda$, and $\forall x_{r,c} \in \mathcal{S}$,

$$P(X_{r,c} = x_{r,c}) = P\left(X_{\tau(r,c)} = x_{r,c}\right) \; ; \; \forall (r,c) \in \Lambda \tag{5.7}$$

The Markov field is said to be homogeneous if and only if the local characteristic (5.5) of the field is site-invariant *i.e.,*

$$\begin{aligned} &P(X_{r,c} = x_{r,c} \,|\, X_{m,n} = x_{m,n} \,;\, (m,n) \in \mathcal{N}(r,c)) = \\ &\qquad P(X_{r',c'} = x_{r',c'} \,|\, X_{m,n} = x_{m,n} \,;\, (m,n) \in \mathcal{N}(r',c')) \;;\; \forall (r,c), (r',c') \in \Lambda \end{aligned} \tag{5.8}$$

Note that, as a consequence of the stationarity property, the measure of the field is space invariant. Hence, the stationarity property implies the homogeneity property (the vice versa is not true). △

Property 5.2 *The marginal probability of each row and column of a Pickard random field presents the structure of a reversible homogeneous Markov chain i.e.,*

$$\boldsymbol{X}_{r,1}^{r,C} \,;\, r \in \{1, \ldots, N\}$$

and

$$\boldsymbol{X}_{1,c}^{R,c} \,;\, c \in \{1, \ldots, N\}$$

are reversible homogeneous Markov chains. All these chains have common initial probabilities $p_i \,;\, i \in \{1, \ldots, N\}$ *and common transition probabilities* $P_{ij} \,;\, i,j \in \{1, \ldots, N\}$*, which can easily be deduced from the measure* τ*.*

This last property justifies our interest in Pickard random field, as discussed in the next chapter. Figure 5.3 provides a graphical representation of the property.

We consider as parameters of the hidden process, the parameters of the Markov chains of rows and columns. Moreover, we adopt the telegraphic parameterization for each chain. Definitions and notations follow from Chapter 2.

Finally, our hidden Markov model based on Pickard random fields can be described

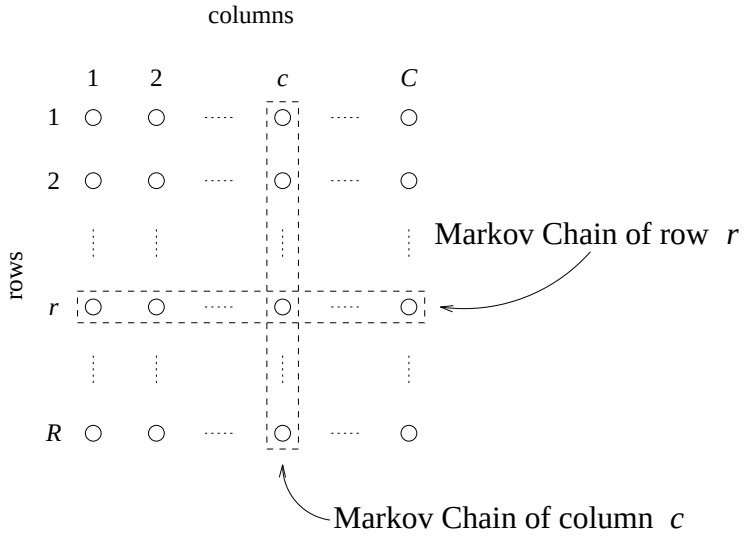

Figure 5.3: Markov Structure of Rows and Columns

by the equations

$$
\begin{aligned}
f\left(\boldsymbol{Y} \mid \boldsymbol{x}\right) &= \prod_{r=1}^{L} \prod_{c=1}^{C} \mathcal{G}_{m_{x_{r,c}}, \sigma^2_{x_{r,c}}}\left(Y_{r,c}\right) \\
P\left(\boldsymbol{x}_{r,1}^{r,C} ; \boldsymbol{\theta}_X\right) &= \frac{\mu_{x_{r,1}} / \left(1-\lambda_{x_{r,1}}\right)}{\sum_i \mu_i / \left(1-\lambda_i\right)} \prod_{c=2}^{C} \left\{\left(1-\lambda_{x_{r,c-1}}\right) \mu_{x_{r,c}} + \delta_{x_{r,c-1} x_{r,c}} \lambda_{x_{r,c}}\right\} \; ; \; r = 1 \ldots R \\
P\left(\boldsymbol{x}_{1,c}^{R,c} ; \boldsymbol{\theta}_X\right) &= \frac{\mu_{x_{1,c}} / \left(1-\lambda_{x_{1,c}}\right)}{\sum_i \mu_i / \left(1-\lambda_i\right)} \prod_{r=2}^{R} \left\{\left(1-\lambda_{x_{r-1,c}}\right) \mu_{r,c} + \delta_{x_{r-1,c} x_{r,c}} \lambda_{x_{r,c}}\right\} \; ; \; c = 1 \ldots C
\end{aligned}
\tag{5.9}
$$

and the $4N$ parameters

$$
\begin{aligned}
\boldsymbol{\theta}_X &= \{\boldsymbol{\mu}, \boldsymbol{\lambda}\} \\
\boldsymbol{\theta}_{Y \mid X} &= \left\{\boldsymbol{m}, \boldsymbol{\sigma}^2\right\}
\end{aligned}
\tag{5.10}
$$

Chapter 6

Maximum Likelihood for Bidimensional HMM Parameters Estimation

6.1 Introduction

In this chapter we consider the estimation of the parameters of the bidimensional hidden Markov model, on the basis of the observations (the realizations of the observable process). As in the unidimensional case, we base our approach on the maximum likelihood technique.

We refer to the model introduced in the previous chapter, where the hidden process is modeled as a Pickard random field. Thus, the parameters to be estimated are the parameters of the observable characteristic $\boldsymbol{\theta}_{Y \mid X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ and the telegraphic parameters of the Markov chains of rows and columns $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$.

According to [Devijver and Dekessel, 1988], we consider an approximation of the likelihood function that allows to exploit the attractive Markovian property of the hidden process rows $\boldsymbol{X}_{r,1}^{r,C}$; $r \in \{1, \ldots, N\}$ and columns $\boldsymbol{X}_{1,c}^{R,c}$; $c \in \{1, \ldots, N\}$ in the model parameter estimation context.

We suppose that rows $\boldsymbol{Y}_{r,1}^{r,C}$; $r \in \{1, \ldots, N\}$ and columns $\boldsymbol{Y}_{r,1}^{r,C}$; $r \in \{1, \ldots, N\}$ of the observable process are mutually independent. Thus, the likelihood function can be approximated by

$$f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) \cong \tilde{f}\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = \prod_{r=1}^{R} \prod_{c=1}^{C} f\left(\boldsymbol{y}_{r,1}^{r,C}\,;\,\boldsymbol{\theta}\right) f\left(\boldsymbol{y}_{1,c}^{R,c}\,;\,\boldsymbol{\theta}\right) \tag{6.1}$$

Note that the assumption of mutually independence is obsviosly not true. Moreover, this approximated likelihood function does not provide the unidimensional likelihood function as a particular case when $R = 1$ or $C = 1$.

Application of Bayes rule to the approximated function allows us to express the

latter as a function of the hidden Markov model components

$$\tilde{f}\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = \prod_{r=1}^{R} f\left(\boldsymbol{y}_{r,1}^{r,C} \mid \boldsymbol{x}_{r,1}^{r,C}\,;\,\boldsymbol{\theta}_{Y\mid X}\right) P\left(\boldsymbol{x}_{r,1}^{r,C}\,;\,\boldsymbol{\theta}_{X}\right) \prod_{c=1}^{C} f\left(\boldsymbol{y}_{1,c}^{R,c} \mid \boldsymbol{x}_{1,c}^{R,c}\,;\,\boldsymbol{\theta}_{Y\mid X}\right) P\left(\boldsymbol{x}_{1,c}^{R,c}\,;\,\boldsymbol{\theta}_{X}\right) \quad (6.2)$$

By definition, the independent terms of the observable characteristic have common parameters, and from Property 5.2 the Markov chains $\boldsymbol{X}_{r,1}^{r,C}\,;\, r \in \{1,\ldots,N\}$ and $\boldsymbol{X}_{1,c}^{R,c}\,;\, c \in \{1,\ldots,N\}$ are also characterized by common parameters. Thus, by taking the above expression into account, the rows $\boldsymbol{Y}_{r,1}^{r,C}\,;\, r \in \{1,\ldots,R\}$ and columns $\boldsymbol{Y}_{1,c}^{R,c}\,;\, c \in \{1,\ldots,C\}$ of the observable process can be considered to be *iid* random vectors, and the approximated likelihood function can be written as

$$\tilde{f}\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = \prod_{l=1}^{R+C} f\left(\boldsymbol{z}_l\,;\,\boldsymbol{\theta}\right) = \prod_{l=1}^{R+C} f\left(\boldsymbol{z}_l \mid \boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\mid X}\right) P\left(\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{X}\right) \quad (6.3)$$

where $\boldsymbol{z}_l$ and $\boldsymbol{v}_l$ are the realizations of the random vectors

$$\boldsymbol{Z}_l = \begin{cases} \boldsymbol{Y}_{l,1}^{l,C} & \text{if } \; 1 \leq l \leq R \\ \boldsymbol{Y}_{1,l}^{R,l} & \text{if } \; R+1 \leq l \leq R+C \end{cases} \quad (6.4)$$

and

$$\boldsymbol{V}_l = \begin{cases} \boldsymbol{X}_{l,1}^{l,C} & \text{if } \; 1 \leq l \leq R \\ \boldsymbol{X}_{1,l}^{R,l} & \text{if } \; R+1 \leq l \leq R+C \end{cases} \quad (6.5)$$

Note that the notation $\boldsymbol{Z}_l$ or $\boldsymbol{V}_l$ represents a vector, with length R or C according to the current value of l.

The *iid* characteristic of the rows and columns of the observable process is equivalent to consider that the observable process is decomposed as in figure 6.1 (in the left we have represented the original process, and in the right we have represented the process "decomposed" in rows and columns), where each element of the decomposition can be treated in the same manner. Note that the values of R and C are in general not equal, hence the "decomposed" structure is not rectangular. Each row $\boldsymbol{Z}_l\,;\, i \in \{1,\ldots,N\}$ of the "decomposed"process may be associated to a hidden Markov chain.

We can already intuitively understand the unidimensional approach to parameter

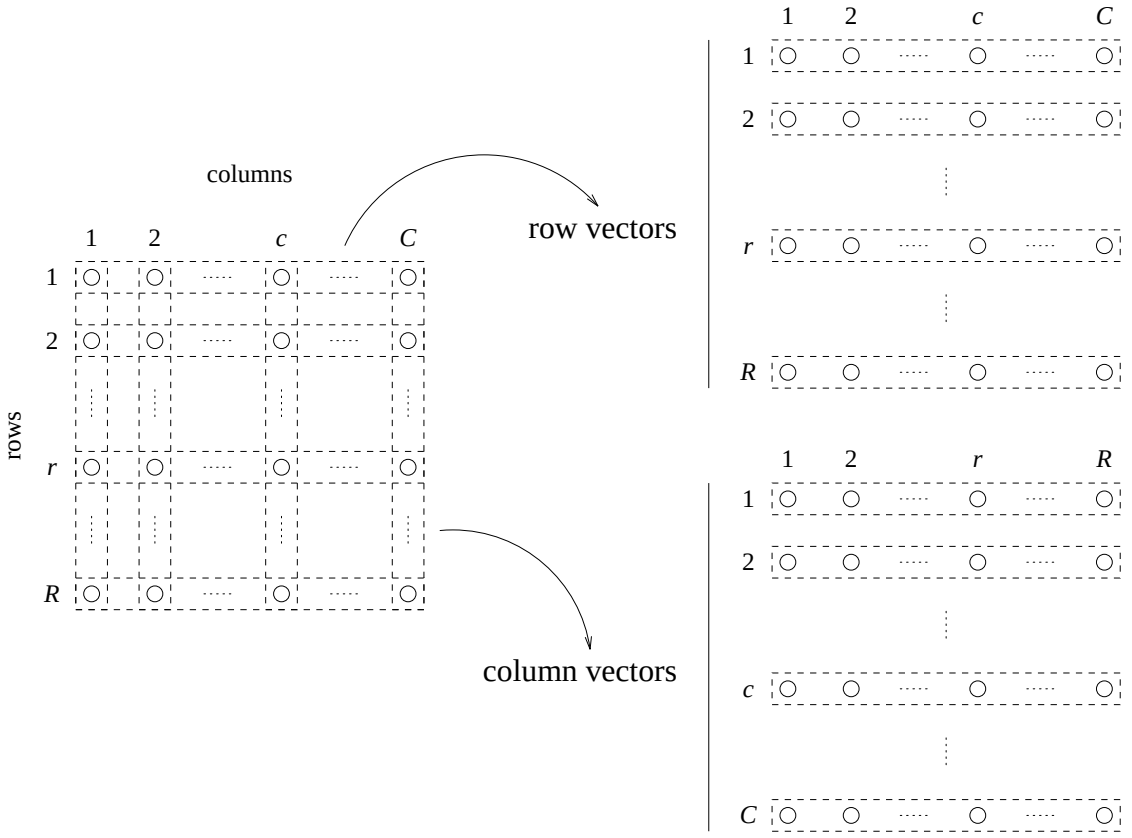

Figure 6.1: Decomposition of the Observable Process

estimation of bidimensional hidden Markov models based on Pickard random fields. Such an approach becomes clear at the approximated likelihood maximization stage.

6.2 Extension of the EM Algorithm to Bidimensional HMM

In Section 3.4 we have introduced the EM algorithm for the maximization of the likelihood function of an hidden Markov chain. Such a maximization was performed by mean of the maximization of an auxiliary function defined in (3.34) and denoted as $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$.

The key of the extension of the EM algorithm to the maximization of the approximated likelihood function resides in the *iid* property of the $\boldsymbol{Z}_l$ random vectors (6.4), as pointed out in [Goussard *et al.*, 1997]. If we substitute the approximated likelihood

function (6.3) in the definition of the auxiliary function (3.34) we obtain

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^0;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\right)=E\left[\ln\prod_{l=1}^{R+C}\tilde{f}\left(\boldsymbol{z}_l\,|\,\boldsymbol{V}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right)P\left(\boldsymbol{V}_l\,;\,\boldsymbol{\theta}_X\right)\;\Big|\;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\,;\,\boldsymbol{\theta}^0\right]$$

$$=\sum_{\boldsymbol{v}_1\ldots\boldsymbol{v}_{R+C}}P\left(\boldsymbol{v}_1\ldots\boldsymbol{v}_{R+C}\,|\,\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\,;\,\boldsymbol{\theta}^0\right)\sum_{l=1}^{R+C}\left\{\ln f\left(\boldsymbol{z}_l\,|\,\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right)+\ln P\left(\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_X\right)\right\}$$

$$=\sum_{\boldsymbol{v}_1\ldots\boldsymbol{v}_{R+C}}P\left(\boldsymbol{v}_1\ldots\boldsymbol{v}_{R+C}\,;\,\boldsymbol{\theta}_X^0\right)\sum_{l=1}^{R+C}\frac{f\left(\boldsymbol{z}_l\,|\,\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}^0\right)}{f\left(\boldsymbol{z}_l\boldsymbol{\theta}^0\right)}\left\{\ln f\left(\boldsymbol{z}_l\,|\,\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right)+\ln P\left(\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_X\right)\right\}$$

$$=\sum_{l=1}^{R+C}\sum_{\boldsymbol{v}_l}P\left(\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_X^0\right)\frac{f\left(\boldsymbol{z}_l\,|\,\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}^0\right)}{f\left(\boldsymbol{z}_l\boldsymbol{\theta}^0\right)}\left\{\ln f\left(\boldsymbol{z}_l\,|\,\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right)+\ln P\left(\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_X\right)\right\}\quad(6.6)$$

The function $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^0;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\right)$ is again composed by two mutually independent terms, which are given by

$$Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X},\boldsymbol{\theta}^0;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\right)=\sum_{l=1}^{R+C}\sum_{\boldsymbol{v}_l}P\left(\boldsymbol{v}_l\,|\,\boldsymbol{z}_l\,;\,\boldsymbol{\theta}^0\right)\ln f\left(\boldsymbol{z}_l\,|\,\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right)\qquad(6.7)$$

and

$$Q_X\left(\boldsymbol{\theta}_X,\boldsymbol{\theta}^0;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\right)=\sum_{l=1}^{R+C}\sum_{\boldsymbol{v}_l}P\left(\boldsymbol{v}_l\,|\,\boldsymbol{z}_l\,;\,\boldsymbol{\theta}_X^0\right)\ln P\left(\boldsymbol{v}_l\,;\,\boldsymbol{\theta}_X\right)\qquad(6.8)$$

Moreover, from (6.3) each term is a summation on l of independent random quantities.

Maximization of $Q_{Y\,|\,X}\left(\boldsymbol{\theta}_{Y\,|\,X},\boldsymbol{\theta}^0;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\right)$ and $Q_X\left(\boldsymbol{\theta}_X,\boldsymbol{\theta}^0;\boldsymbol{z}_1\ldots\boldsymbol{z}_{R+C}\right)$ can be performed as in the unidimensional case: we need only to take the summation over the index l into account.

6.2.1 Re-estimation Transformation for $\boldsymbol{\theta}_{Y \mid X}$

From Section 3.5, and by considering the definitions (6.4) and (6.5), the maximization of (6.7) with respect to the parameters $\boldsymbol{\theta}_{Y \mid X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ yields

$$m_i = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{P\left(X_{r,c}=i \mid \boldsymbol{y}_{r,1}^{r,C};\boldsymbol{\theta}^0\right)+P\left(X_{r,c}=i \mid \boldsymbol{y}_{1,c}^{R,c};\boldsymbol{\theta}^0\right)\right\} y_{r,c}}{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{P\left(X_{r,c}=i \mid \boldsymbol{y}_{r,1}^{r,C};\boldsymbol{\theta}^0\right)+P\left(X_{r,c}=i \mid \boldsymbol{y}_{1,c}^{R,c};\boldsymbol{\theta}^0\right)\right\}}$$

$$\sigma_i^2 = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{P\left(X_{r,c}=i \mid \boldsymbol{y}_{r,1}^{r,C};\boldsymbol{\theta}^0\right)+P\left(X_{r,c}=i \mid \boldsymbol{y}_{1,c}^{R,c};\boldsymbol{\theta}^0\right)\right\} (y_{r,c}-m_i)^2}{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{P\left(X_{r,c}=i \mid \boldsymbol{y}_{r,1}^{r,C};\boldsymbol{\theta}^0\right)+P\left(X_{r,c}=i \mid \boldsymbol{y}_{1,c}^{R,c};\boldsymbol{\theta}^0\right)\right\}}$$

Let us introduce the following notation

$$\begin{aligned}
&\overset{c\mid}{\mathcal{F}}_r(i) \;:\; \text{element } r \text{ of the Forward variable associated to the MC of column } c\\
&\overset{r-}{\mathcal{F}}_c(i) \;:\; \text{element } c \text{ of the Forward variable associated to the MC of row } r\\
&\overset{c\mid}{\mathcal{B}}_r(i) \;:\; \text{element } r \text{ of the Backward variable associated to the MC of column } c\\
&\overset{r-}{\mathcal{B}}_c(i) \;:\; \text{element } c \text{ of the Backward variable associated to the MC of row } r
\end{aligned} \tag{6.9}$$

By taking into account the relation (3.38) between the probabilities $P\left(X_{r,c}=i \mid \boldsymbol{y}_{*,*}^{*,*};\boldsymbol{\theta}^0\right)$ and Forward and Backward variables, the re-estimation equations of the parameters

$\boldsymbol{\theta}_{Y\ X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ may be written as

$$m_i = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{ \overset{r-}{\mathcal{F}^0}_c(i)\,\overset{r-}{\mathcal{B}^0}_c(i) + \overset{c\ |}{\mathcal{F}^0}_r(i)\,\overset{c\ |}{\mathcal{B}^0}_r(i) \right\} y_{r,c}}{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{ \overset{r-}{\mathcal{F}^0}_c(i)\,\overset{r-}{\mathcal{B}^0}_c(i) + \overset{c\ |}{\mathcal{F}^0}_r(i)\,\overset{c\ |}{\mathcal{B}^0}_r(i) \right\}} \tag{6.10}$$

$$\sigma_i^2 = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{ \overset{r-}{\mathcal{F}^0}_c(i)\,\overset{r-}{\mathcal{B}^0}_c(i) + \overset{c\ |}{\mathcal{F}^0}_r(i)\,\overset{c\ |}{\mathcal{B}^0}_r(i) \right\} (y_{r,c} - m_i)^2}{\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{ \overset{r-}{\mathcal{F}^0}_c(i)\,\overset{r-}{\mathcal{B}^0}_c(i) + \overset{c\ |}{\mathcal{F}^0}_r(i)\,\overset{c\ |}{\mathcal{B}^0}_r(i) \right\}} \tag{6.11}$$

The re-estimates of the parameters are then obtained by "accumulation" of unidimensional quantities.

From the definition (6.3) of the approximated likelihood function, each marginal distribution in the product has the form of the likelihood function associated to an hidden Markov model. Thus, each one of them suffers of the degeneracy problem discussed in Section 3.2. By adopting the Bayesian solution proposed in Section 3.2.1, from the penalized re-estimation formula derived in the unidimensional case (3.52), the re-estimation formula (6.11) reads

$$\sigma_i^2 = \frac{\alpha + \frac{1}{2}\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{ \overset{r-}{\mathcal{F}^0}_c(i)\,\overset{r-}{\mathcal{B}^0}_c(i) + \overset{c\ |}{\mathcal{F}^0}_r(i)\,\overset{c\ |}{\mathcal{B}^0}_r(i) \right\} (y_{r,c} - m_i)^2}{\beta + \frac{1}{2}\sum_{r=1}^{R}\sum_{c=1}^{C}\left\{ \overset{r-}{\mathcal{F}^0}_c(i)\,\overset{r-}{\mathcal{B}^0}_c(i) + \overset{c\ |}{\mathcal{F}^0}_r(i)\,\overset{c\ |}{\mathcal{B}^0}_r(i) \right\}} \tag{6.12}$$

Akin to the penalized re-estimation formula in the unidimensional case (3.52), the above re-estimation formula gives σ^2 estimates strictly greater than zero.

6.2.2 Re-estimation Transformation for $\boldsymbol{\theta}_X$

In order to estimate the telegraphic parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ we refer to the telegraphic EM algorithm (TEM) introduced in Chapter 4.

From the definitions of the function $Q_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ with telegraphic parameters (3.46) and the function $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ of the telegraphic EM algorithm (4.4-4.7), that we have given in the unidimensional case, the bidimensional function $R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right)$ can

be defined as

$$
\begin{aligned}
&R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = \\
&\sum_{j=1}^{N}\sum_{c=1}^{C}\frac{1}{2}\left\{\sum_{r=1}^{R} P\left(X_{r,c}=j \mid \boldsymbol{y}_{1,c}^{R,c};\, \boldsymbol{\theta}^0\right) + \sum_{r=2}^{R-1} P\left(X_{r,c}=j \mid \boldsymbol{y}_{1,c}^{R,c};\, \boldsymbol{\theta}^0\right)\right\} \ln \mu_j \left(1-\lambda_j\right) \\
&\qquad + \sum_{j=1}^{N}\sum_{c=1}^{C}\sum_{r=2}^{R} P\left(X_{r-1,c}=j, X_{r-1,c}=j \mid \boldsymbol{y}_{1,c}^{R,c};\, \boldsymbol{\theta}^0\right) \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} \\
&+ \sum_{j=1}^{N}\sum_{r=1}^{R}\frac{1}{2}\left\{\sum_{c=1}^{C} P\left(X_{r,c}=j \mid \boldsymbol{y}_{r,1}^{r,C};\, \boldsymbol{\theta}^0\right) + \sum_{c=2}^{C-1} P\left(X_{r,c}=j \mid \boldsymbol{y}_{r,1}^{r,C};\, \boldsymbol{\theta}^0\right)\right\} \ln \mu_j \left(1-\lambda_j\right) \\
&\qquad + \sum_{j=1}^{N}\sum_{r=1}^{R}\sum_{c=2}^{C} P\left(X_{r,c-1}=j, X_{r,c-1}=j \mid \boldsymbol{y}_{r,1}^{r,C};\, \boldsymbol{\theta}^0\right) \ln \frac{\left(1-\lambda_j\right)\mu_j + \lambda_j}{\left(1-\lambda_j\right)\mu_j} \qquad (6.13)
\end{aligned}
$$

By defining

$$\overset{\perp}{\alpha}{}_i^0 = \sum_{c=1}^{C}\sum_{r=1}^{R}\left\{P\left(X_{r,c}=j \mid \boldsymbol{y}_{1,c}^{R,c};\, \boldsymbol{\theta}^0\right) + P\left(X_{r,c}=j \mid \boldsymbol{y}_{r,1}^{r,C};\, \boldsymbol{\theta}^0\right)\right\} \qquad (6.14)$$

$$\overset{\perp}{\beta}{}_i^0 = \sum_{c=1}^{C}\sum_{r=2}^{R-1} P\left(X_{r,c}=i \mid \boldsymbol{y}_{1,c}^{R,c};\, \boldsymbol{\theta}^0\right) + \sum_{r=1}^{R}\sum_{c=2}^{C-1} P\left(X_{r,c}=i \mid \boldsymbol{y}_{r,1}^{r,C};\, \boldsymbol{\theta}^0\right) \qquad (6.15)$$

$$\overset{\perp}{s}{}_i^0 = \sum_{c=1}^{C}\sum_{r=2}^{R} P\left(X_{r-1,c}=i, X_{r-1,c}=i \mid \boldsymbol{y}_{1,c}^{R,c};\, \boldsymbol{\theta}^0\right) \qquad (6.16)$$

$$+ \sum_{r=1}^{R}\sum_{c=2}^{C} P\left(X_{r,c-1}=i, X_{r,c-1}=i \mid \boldsymbol{y}_{r,1}^{r,C};\, \boldsymbol{\theta}^0\right) \qquad (6.17)$$

and

$$\overset{\perp}{\gamma}{}_i^0 = \frac{\overset{\perp}{\alpha}{}_i^0 + \overset{\perp}{\beta}{}_i^0}{2} \qquad (6.18)$$

the function reads

$$R_X\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0; \boldsymbol{y}\right) = \sum_{j=1}^{N} \overset{\perp}{\gamma}{}_i^0 \ln \mu_j \left(1-\lambda_j\right) + \overset{\perp}{s}{}_i^0 \ln\left(\left(1-\lambda_j\right)\mu_j + \lambda_j\right) \qquad (6.19)$$

The re-estimation of the telegraphic parameter is done exactly as in the unidimensional case. The bidimensional re-estimation formulas are obtained by substituting the quantities s_i^0 and γ_i^0 with $\overset{\perp}{s_i^0}$ and $\overset{\perp}{\gamma_i^0}$ in the unidimensional formulas. The symbol $\perp$ denotes the dependence of the bidimensional quantities on both rows and colums.

Again we obtain the parameter estimates by "accumulation" of unidimensional quantities, where the accumulation resides in $\overset{\perp}{s_i^0}$ and $\overset{\perp}{\gamma_i^0}$.

Part II

Medical Image Segmentation

Chapter 7

Medical Image Segmentation

7.1 Introduction

As discussed in the introduction, the segmentation operation provides an efficient method to distinguish different uniform zones of the image. In our applied issue, this means to allow the computation of the inner areas of the ventricle.

Such an operation consists in dividing the image in different zones according to a uniformity criterion that has to be defined.

Within the framework of a model-based approach, we must consider a model that implements the uniformity criterion, while allowing the calculation of the segmented image with a reasonable computational burden: the hidden Markov models that we have studied in the first part of this work well fit this purpose.

7.2 Segmentation Model

Hidden Markov models have been defined as the composition of a hidden Markovian process and an observable process, where in the bidimensional case the hidden process is a Markov random field.

Markov random fields well play the role of a uniformity criterion, since they model qualitative information about scenes and textures such as piecewise uniformity in somewhat natural way. As a consequence, they well fit the modeling of the segmented image, since the latter is by definition characterized of uniform zones separated by sharp discontinuities.

In particular, we base the modeling of the segmented image on Pickard random fields, and we consider the parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ of the Markov chain, on each row and column of the random field, as the hidden process parameters. Such a choice will be justified by the method used to obtain the segmented image, as discussed in the following section.

For a N level segmentation, the state space of the samples (realizations) of the

Pickard random field is supposed to have N elements *i.e.,* $|\mathcal{S}| = N$.

The image that has to be segmented (the original image) represents the observations and it is obviously modeled by the observable process. Such an image can be considered as a noise corrupted version of the segmented image: the noise introduces a certain smoothness or blurring and can be somehow considered as the inverted operation of the segmentation. Thus, the relation between the observable process $\boldsymbol{Y} = \{Y_{r,c}\}_{(r,c)\in\Lambda}$ and the hidden process $\boldsymbol{X} = \{X_{r,c}\}_{(r,c)\in\Lambda}$ may be defined as

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{B} \tag{7.1}$$

where $\boldsymbol{B} = \{B_{r,c}\}_{(r,c)\in\Lambda}$ is the stochastic process that models the noise.

With the assumption that $\boldsymbol{B} = \{B_{r,c}\}_{(r,c)\in\Lambda}$ is a white Gaussian noise (*i.e.,* the random variables $B_{r,c_{(r,c)\in\Lambda}}$ are independent with a Gaussian distribution), from the equivalence

$$f\left(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{X} = \boldsymbol{x}\right) = f\left(\boldsymbol{B} + \boldsymbol{X} = \boldsymbol{y} \,|\, \boldsymbol{X} = \boldsymbol{x}\right) = f\left(\boldsymbol{B} + \boldsymbol{x} = \boldsymbol{y}\right) = f\left(\boldsymbol{B} = \boldsymbol{y} - \boldsymbol{x}\right)$$

the observable characteristic is the product of independent Gaussian distributions. The mean values $\boldsymbol{m}$ of the Gaussian distribution are specified by the values of the states (the possible realizations of each $X_{r,c}\,;\,(r,c) \in \Lambda$), and we suppose to have a different variance for each different mean value. Thus, the observable characteristic depends on the parameters $\boldsymbol{\theta}_{Y\,|\,X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$, and corresponds to the definition given in (2.8).

With these specifications, the model for the image segmentation is the one we have defined in Chapter 5.

7.3 Segmented Image Computation

We consider as the segmented image the most probable sample of the hidden process. To obtain such a realization we maximize the marginal *a posteriori* distributions of the Pickard random field (the hidden process) *i.e.,* we compute

$$x_{\mathrm{seg}_{r,c}} = \underset{i\in\{1,\ldots,N\}}{\arg\max}\, P\left(X_{r,c} = i \,|\, \boldsymbol{Y} = \boldsymbol{y}\,;\,\boldsymbol{\theta}\right)\,;\,(r,c) \in \Lambda \tag{7.2}$$

Each $P\left(X_{r,c} \,|\, \boldsymbol{Y} = \boldsymbol{y}\,;\,\boldsymbol{\theta}\right)$ is called the *a posteriori* marginal distribution of the site $(r,c) \in \Lambda$ ($\boldsymbol{\theta}$ are the parameters of the model).

The estimation of the realization through the maximization of the *a posteriori* distribution (or the marginal *a posteriori* distributions) is known as *maximum a posteriori* (MAP) *likelihood method* [Demoment, 1996]. The *a posteriori* attribute is due to the fact that one considers the distribution of the random field given the observations and

the parameters of the model.

According to [Devijver and Dekessel, 1988], each marginal *a posteriori* distribution is approximated by assuming that the dependence on the observable process $\boldsymbol{Y} = \{Y_{r,c}\}_{(r,c)\in\Lambda}$ is limited to $\boldsymbol{Y}_{r,1}^{r,C}$ and $\boldsymbol{Y}_{1,c}^{R,c}$ (the row r and the column c of the observable process, respectively):

$$P\left(X_{r,c} \,|\, \boldsymbol{Y} = \boldsymbol{y}\,;\, \boldsymbol{\theta}\right) \cong P\left(X_{r,c} \,|\, \boldsymbol{Y}_{r,1}^{r,C} = \boldsymbol{y}_{r,1}^{r,C}, \boldsymbol{Y}_{1,c}^{R,c} = \boldsymbol{y}_{1,c}^{R,c}\,;\, \boldsymbol{\theta}\right) \tag{7.3}$$

This assumption means that the *a posteriori* distribution takes into account the realizations of the observable process only on a cross shaped set of sites instead of all the sites, as represented in figure 7.1.

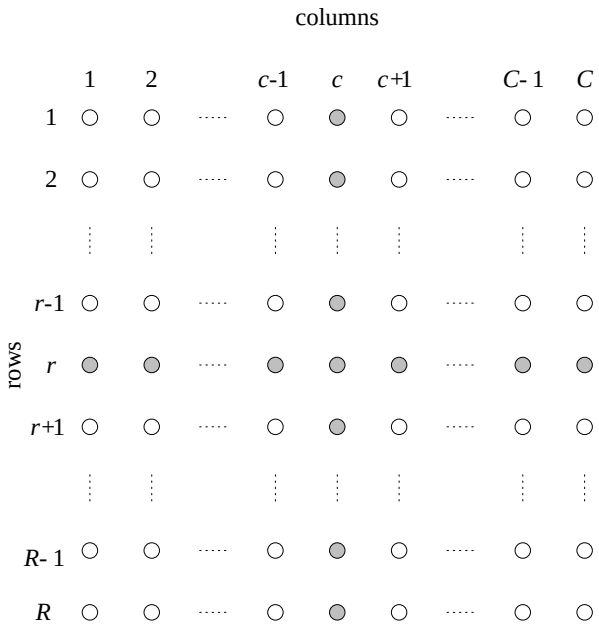

Figure 7.1: Dependence on a Cross Shaped Set of Sites

The benefit from such an approximation is to allow the expression of the marginal *a posteriori* distribution as a function of the Forward and Backward variables by rows and columns defined in (6.9) ($\overset{r}{\overline{\mathcal{F}}}_c(i)$, $\overset{r}{\overline{\mathcal{B}}}_c(i)$, $\overset{c}{\underline{\mathcal{F}}}_r(i)$, $\overset{c}{\underline{\mathcal{B}}}_r(i)$; $(r,c) \in \Lambda$, $i \in \{1,\dots N\}$). Indeed, according to [Devijver and Dekessel, 1988] the approximated marginal *a priori* distribution reads

$$P\left(X_{r,c} = i \,|\, \boldsymbol{Y}_{r,1}^{r,C} = \boldsymbol{y}_{r,1}^{r,C}, \boldsymbol{Y}_{1,c}^{R,c} = \boldsymbol{y}_{1,c}^{R,c}\,;\, \boldsymbol{\theta}\right) \propto \overset{r}{\overline{\mathcal{F}}}_c(i)\, \overset{r}{\overline{\mathcal{B}}}_c(i)\, \overset{c}{\underline{\mathcal{F}}}_r(i)\, \overset{c}{\underline{\mathcal{B}}}_r(i)\,;$$
$$i \in \{1,\dots,N\}, (r,c) \in \Lambda \tag{7.4}$$

The proof of the above relation can be obtained by means of multiple applications of Bayes rule, similarly to the proof of the relations between hidden Markov chain quantities and the Forward and Backward variables that we have developed in Appendix A.

For the maximization purpose, the approximated *a posteriori* distribution is then completely characterized by the Forward and Backward variables by rows and columns. Note that such variables depend on the parameters $\boldsymbol{\theta}_{Y|X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ of the observable characteristic and on the parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ of the Markov chains associated to the hidden process rows and columns. Hence, the bidimensional hidden Markov model that we have introduced in Chapter 5 contains all the parameters that are necessary in order to compute the segmented image.

Once the parameters of the model are estimated, by means of the bidimensional extension of the TEM algorithm of Chapter 6, we can compute the Forward and Backward variables by rows and columns and then we can have access to the approximated marginal *a posteriori* distributions (7.4). The number of possible state values is finite and often small, since it corresponds to the number of segmentation levels. Therefore, the maximization of the approximated *a posteriori* distribution can be performed by a direct evaluation on each possible state *i.e.,*

$$x_{\mathrm{seg}_{r,c}} = \arg\max_{i \in \{1,\dots,N\}} \overset{r}{\mathcal{F}}_c(i)\, \overset{r}{\mathcal{B}}_c(i)\, \overset{c}{\mathcal{F}}_r(i)\, \overset{c}{\mathcal{B}}_r(i) \;;\; (r,c) \in \Lambda \tag{7.5}$$

Note that the computational burden of such a maximum *a posteriori* technique resides in the computation of the Forward and Backward variables.

In Appendix B a scheme of the complete segmentation algorithm is provided.

7.4 Results

In this section we show an example of non convergence of the TEM algorithm, in the case of a unidimensional set of observations (generated by noise corrupted Markov chain samples), and the results of the segmentation of an X-ray tomography image of the heart.

We have considered a nine level segmentation, which provides a good distinction between different uniform zones, and, as an example of excessive segmentation, a three level segmentation. Note that, in the introduction we have intuitively discuss the possible result of a three level segmentation: the actual result should convince the reader that the segmentation operation is not as easy as the one performed intuitively.

We have obtained the results by means of the program described in Appendix B, which performs the maximum likelihood estimation of the parameters and computes the segmented image with a maximum *a posteriori* technique.

For the nine level segmentation, we consider three different set of maximum like-

lihood estimates of the parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$, obtained by means of the TEM algorithm, the mixed TEM - gradient descent algorithm and the gradient descent algorithm, respectively (such algorithms are described in Appendix B).

For each set we show the corresponding segmented image and the maximum values of the marginal *a posteriori* distribution. Moreover, in order to evaluate the three methods, we show the evolution of the values of the negative log likelihood (NLL) and the values of the norm of the gradient with respect to the parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$, during the iterative estimation procedure.

Then, we expose a comparative study of the three methods in terms of the evolution of the values of the negative log likelihood and the values of the gradient norm.

The original image is shown in figure 7.2. Note that rows and columns of the image are oriented according to the lattice described in figure 5.1.

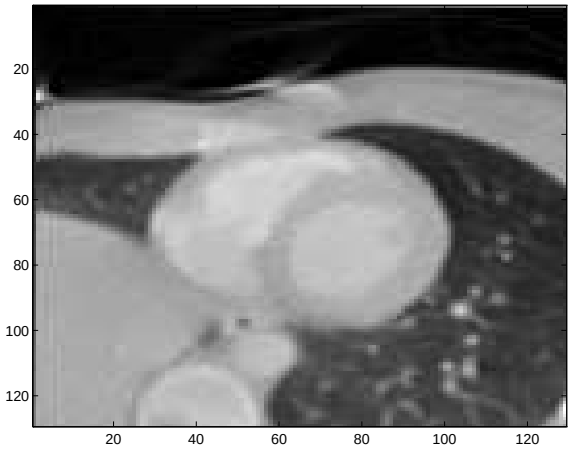

Figure 7.2: Original Image

7.4.1 Example of Non Convergence

The example is characterized by a 16 level segmentation of an observation set composed by 253 samples. Figure 7.3 shows the original signal (above) and the segmented signal (below).

Figure 7.4 shows the evolution of two of the 16 parameters $\boldsymbol{\mu}$ (μ_1 and μ_2). On the left we have the whole domain of such parameters, where we can observe that after an initial evolution (note the difference between the initial values and the first estimate) the values of the two parameters remain within the intervals $\mu_1 \in [0.0370, 0.0375]$, $\mu_2 \in [0.0265, 0.0270]$. The close-up picture on the right shows the oscillation of the parameters within such intervals.

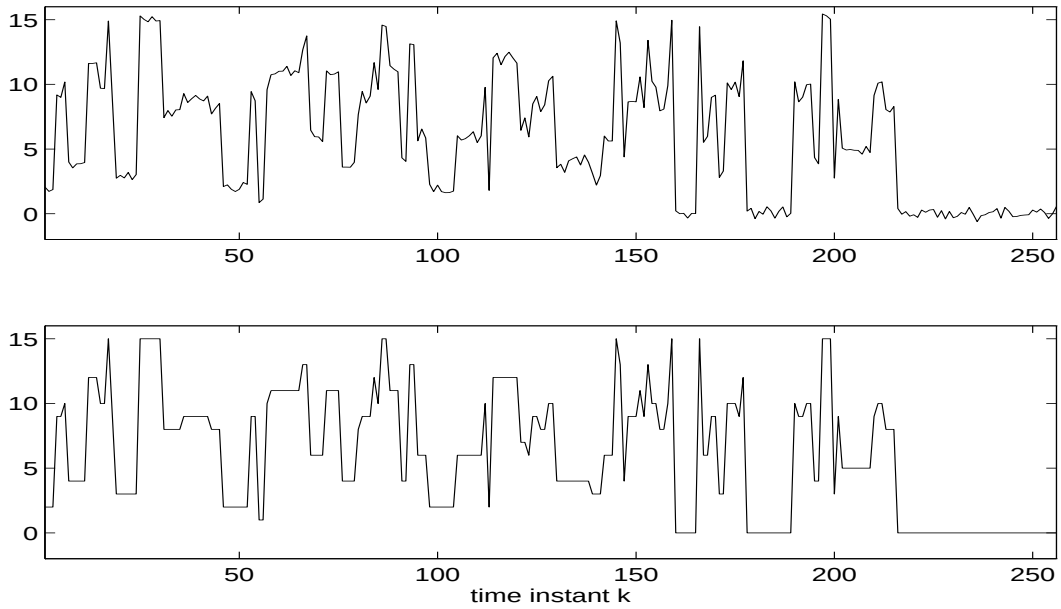

Figure 7.3: Original and Segmented Signal

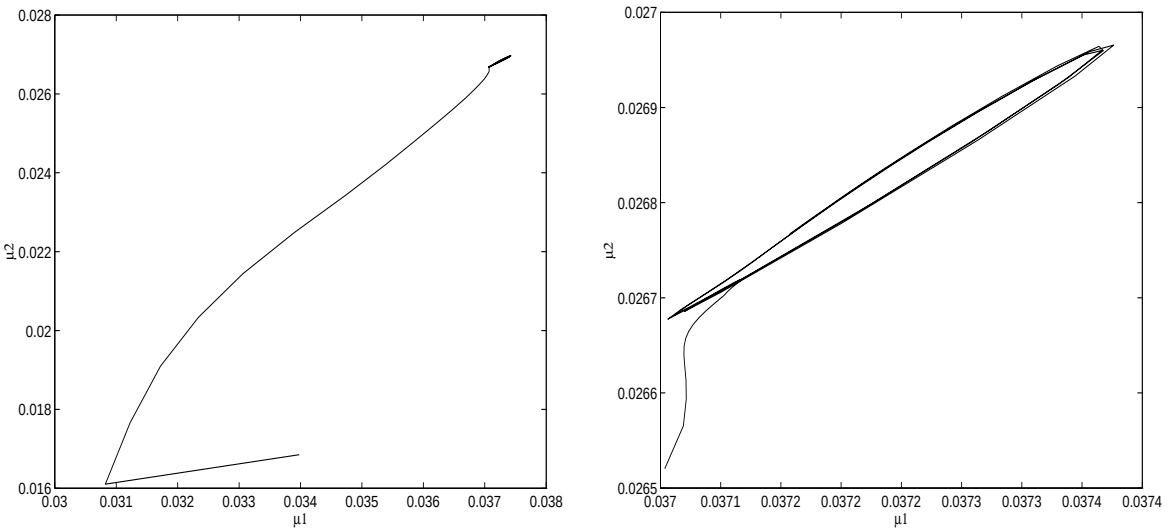

Figure 7.4: Non Convergence of μ_1 and μ_2

The evolution of the negative log likelihood is provided in figure 7.5: after approximately 20 iterations it begins to oscillate. This example clearly shows that the TEM algorithm does not guarantee that the likelihood increases, and therefore does not guarantee the systematic convergence to a local maximum of the likelihood.

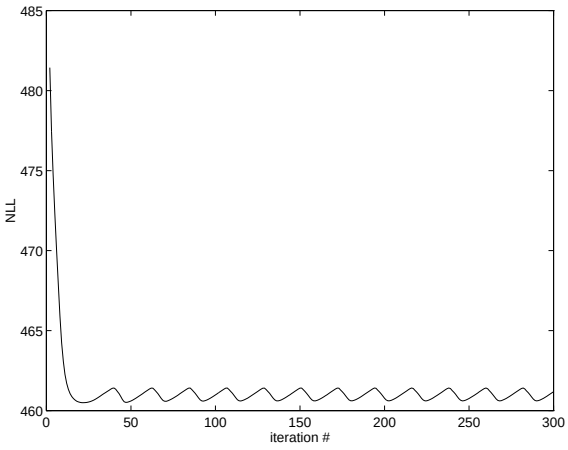

Figure 7.5: Non Convergence of the NLL

7.4.2 Nine Level Segmentation

The segmentation on 9 levels provides a discrete distinction between the uniform areas of the images, as may be observed from the segmented images obtained by estimating the parameters with the three methods.

In the following tables we provide some characteristic values of the three methods: in the first table we provide for each algorithm the number of iterations, the value of the gradient norm and the value of the negative log likelihood once the corresponding iterative procedure has terminated; in the second table we provide the values of the gradient norm and the negative log likelihood at the number of iterations of the iterative procedure that terminates first (in our case the TEM - gradient method terminates first, after 53 iterations).

	TEM	TEM-Gradient	Gradient
number of iterations	75	53	368
final value of $\lvert\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ln f(\boldsymbol{y};\boldsymbol{\theta})\rvert$	$2.7104e+3$	608	133.22
final value of $-\ln f(\boldsymbol{y};\boldsymbol{\theta})$	$1.2284e+5$	$1.2286e+5$	$1.2422e+5$

	TEM	TEM-Gradient	Gradient
value of $\lvert\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ln f(\boldsymbol{y};\boldsymbol{\theta})\rvert$	$2.7101e+3$	608	$1.6022e+3$
value of $-\ln f(\boldsymbol{y};\boldsymbol{\theta})$	$1.2286e+5$	$1.2284e+5$	$1.2438e+5$

7.4.2.1 TEM Algorithm

Let us consider the model parameter estimates obtained by means of the TEM algorithm. Figure 7.6 shows the segmented image (left) and the maximum values of the *a posteriori* distribution of the hidden process.

In the image of the maximum values of the *a posteriori* distribution of the hidden process, the white color denotes the probability value 1 and the black color denotes the probability value 0. It is interesting to observe that within a uniform area there was no incertitude in assigning the level, since the chosen level has a corresponding probability closer to 1, while in the area between two different zones the probability is distributed among different levels.

Figure 7.7 provides the values of the gradient norm an the value of the negative log likelihood. Note that after about 30 iterations the gradient stabilizes to a value different than zero.

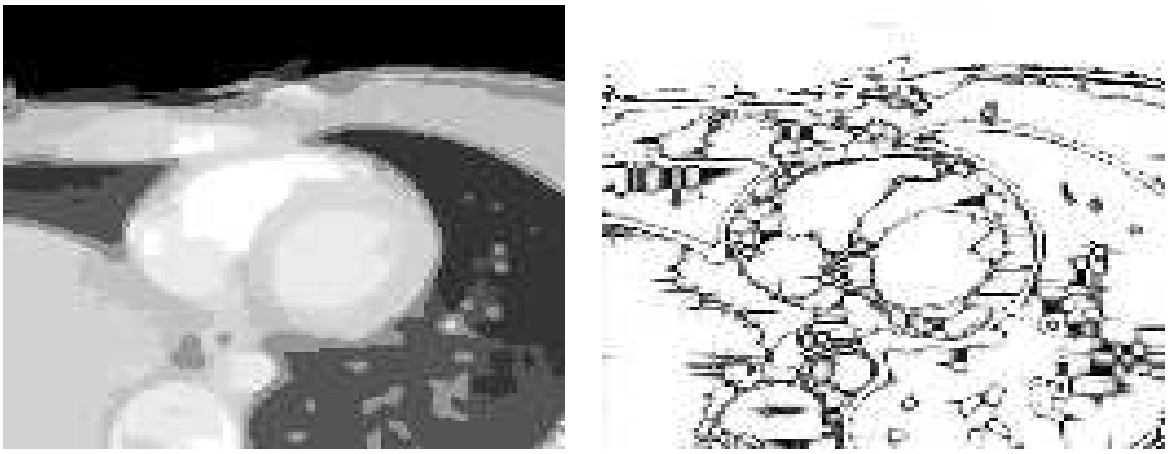

Figure 7.6: TEM 9 Level Segmentation: Segmented Image and MAP Values

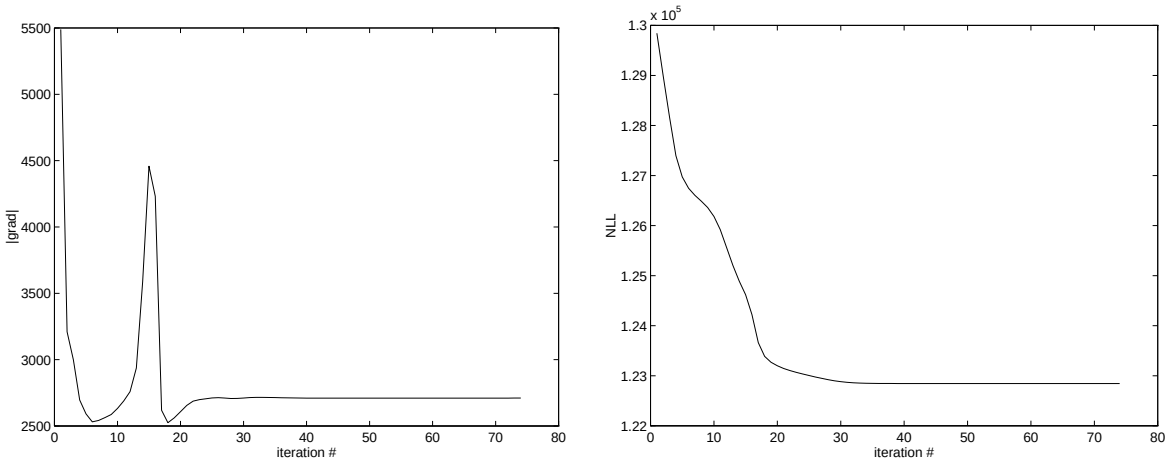

Figure 7.7: TEM 9 Level Segmentation: $|\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ln f(\boldsymbol{y}; \boldsymbol{\theta})|$, $-\ln f(\boldsymbol{y}; \boldsymbol{\theta})$

7.4.2.2 TEM - Gradient Descent Algorithm

Let us consider the model parameter estimates obtained by means of the mixed TEM - gradient descent algorithm. Figure 7.6 shows the segmented image (left) and the maximum values of the *a posteriori* distribution, while figure 7.7 shows the values of the gradient norm an the value of the negative log likelihood.

Although the segmented image does not show evident differences with respect to the outcome obtained with TEM model parameters estimation, we can note that, as soon as we switch from the TEM algorithm to the gradient descent algorithm (iteration # 30), the gradient norm decreases considerably.

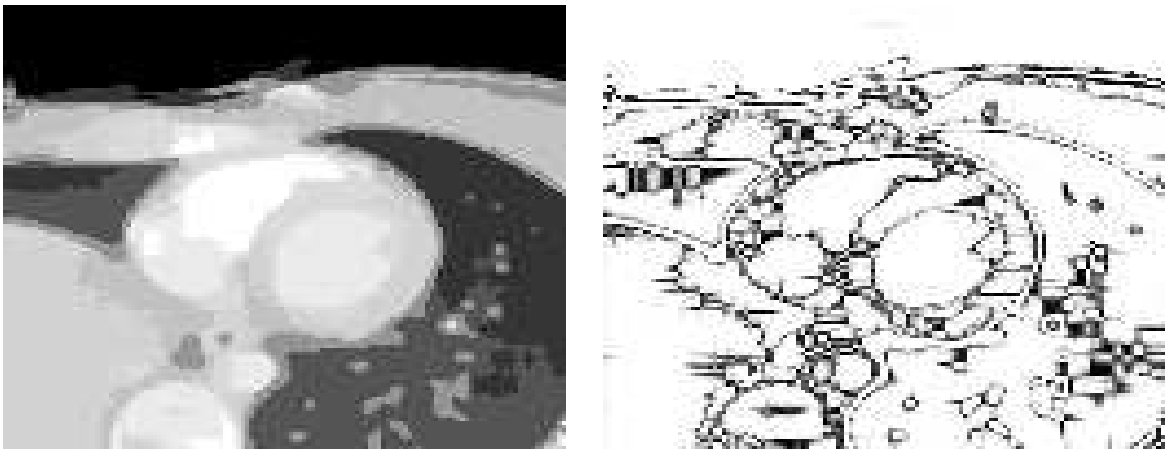

Figure 7.8: TEM-Gradient 9 Level Segmentation: Segmented Image and MAP Values

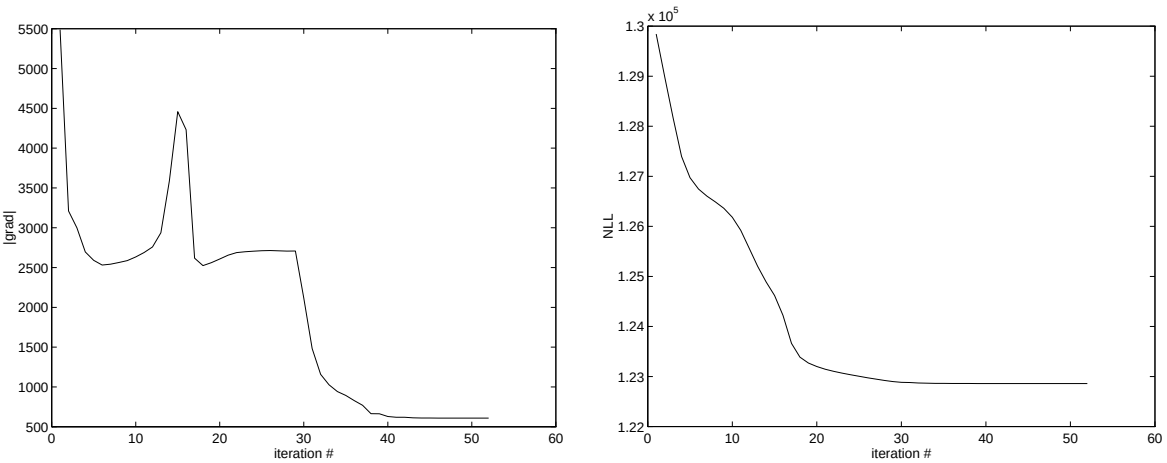

Figure 7.9: TEM-Gradient 9 Level Segmentation: $|\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ln f(\boldsymbol{y};\boldsymbol{\theta})|$, $-\ln f(\boldsymbol{y};\boldsymbol{\theta})$

7.4.2.3 Gradient Descent Algorithm

Let us consider the model parameter estimates obtained by mean a Gradient Descent method, which is initialized on the same point as the TEM algorithm. Figure 7.6 shows the segmented image (left) and the maximum values of the *a posteriori* distribution, while figure 7.7 shows the values of the gradient norm an the value of the negative log likelihood. The segmentation result and the evolution of the negative log likelihood are different from the ones obtained with the previous two methods: from such a difference we deduce that the gradient descent has attained a different local minimum of the negative log likelihood (thus, a different local maximum of the likelihood function). It is interesting to remark the oscillations of the norm of the gradient, which are probably due to the non optimization of the descent step (see Appendix B).

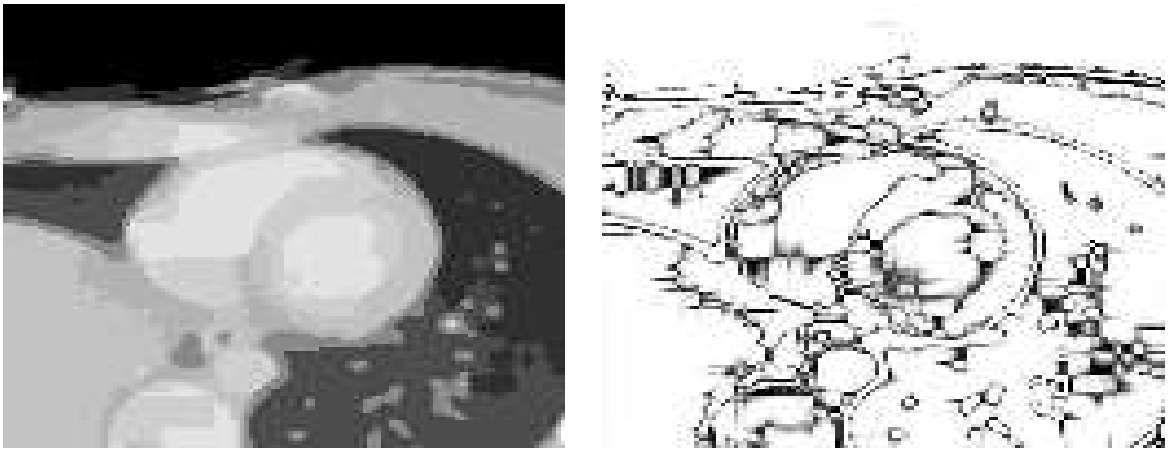

Figure 7.10: Gradient 9 Level Segmentation: Segmented Image and MAP Values

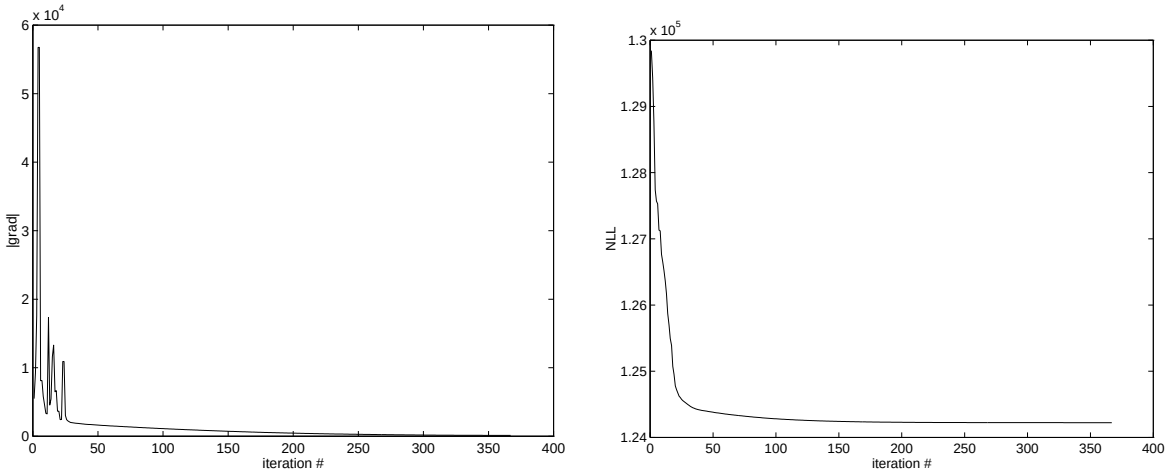

Figure 7.11: Gradient 9 Level Segmentation: $|\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ln f(\boldsymbol{y};\boldsymbol{\theta})|$, $-\ln f(\boldsymbol{y};\boldsymbol{\theta})$

7.4.2.4 Comparison of the Three Methods

In this section we provide a qualitative comparison of the three methods. Figure 7.12 and figure 7.13 shows a close-up picture respectively of the value of the gradient norm and the value of the negative *log* likelihood, around the iteration at which the TEM-gradient descent algorithm switches form the TEM method to the gradient descent method. With a gradient descent method initialized with the TEM estimate at such an iteration, the likelihood function gets closer to a local maximum, and therefore, the parameter estimation is improved.

We must recognize that it is a moderate improvement (the gradient does not tend to zero and the likelihood decreases of a small quantity) but, on the basis of the results we have obtained, we think that with an optimization of the descent step such an improvement should be considerable. Moreover, the mixed method computes the parameter estimates with a smaller number of iterations than the one required by the other two methods.

We can observe that although the gradient descent method is the method that most decreases the gradient value, it has a slow convergence rate and attains a different local minimum (see Section 7.4.2.3).

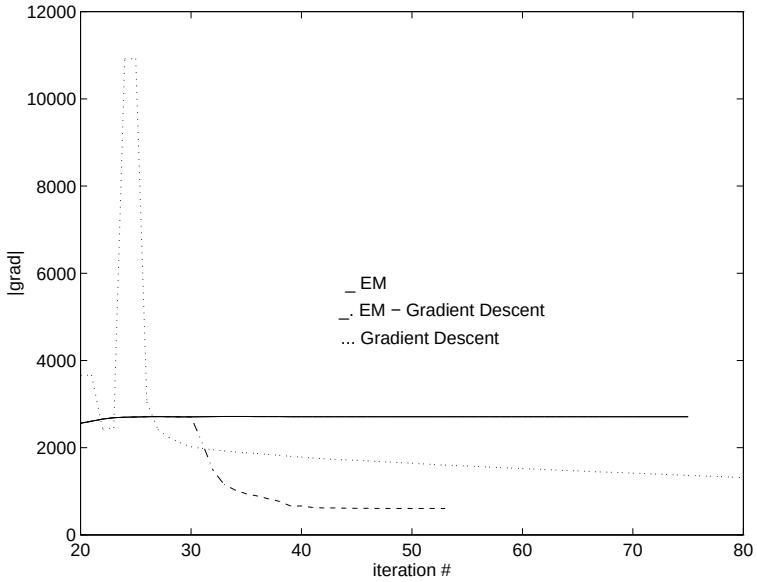

Figure 7.12: Comparison of $|\nabla_{\boldsymbol{\mu},\boldsymbol{\lambda}} \ln f(\boldsymbol{y}; \boldsymbol{\theta})|$

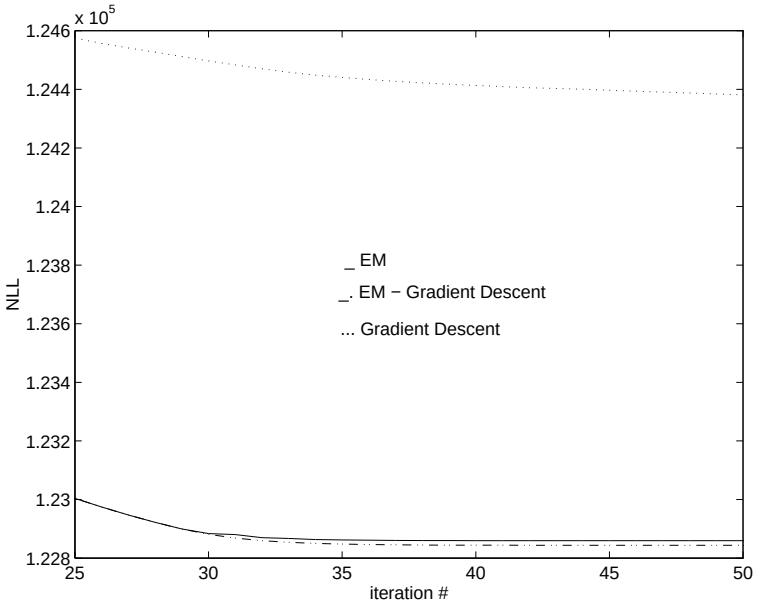

Figure 7.13: Comparison of $-\ln f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right)$

7.4.3 Three Level Segmentation

The image segmented with three level provides an example of excessive segmentation. From the original image we can observe that three levels are not sufficient to represent the different uniform areas: the algorithm "forces the uniformity" and different zones are mixed, as it is clearly shown in figure 7.14.

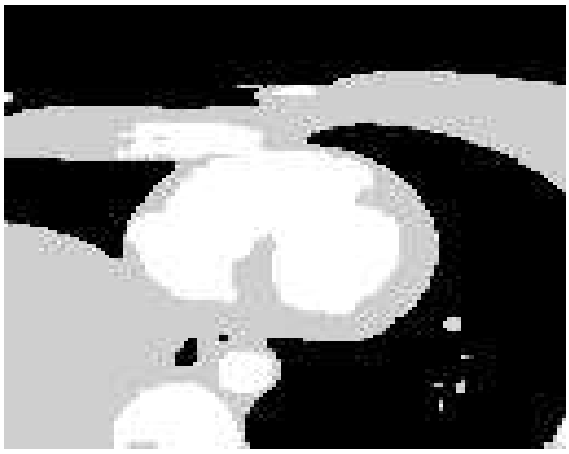

Figure 7.14: TEM 3 Level Segmented Image

Appendix A

The Normalized Forward - Backward algorithm

A.1 Introduction

In Section 3.3 we have introduced the well known *normalized Forward - Backward* algorithm, in order to efficiently evaluate the likelihood function associated to a hidden Markov model. Later on (Section 3.4.2) we have referred again to such an algorithm for the computation of the probabilities $P\left(X_{k-1}=i, X_k=j \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right)$ and $P\left(X_k=i \mid \boldsymbol{y}, \boldsymbol{\theta}^0\right)$.

Let us consider the hidden Markov model defined in Chapter 2 and all the notations associated to such a definition.

The normalized Forward - Backward algorithm is defined by the *Forward variable*

$$\mathcal{F}_k(i) = P\left(X_k = i \mid \boldsymbol{y}_1^k ; \boldsymbol{\theta}\right) \tag{A.1}$$

which is computed with a forward recurrence

$$\begin{aligned} \mathcal{F}_1(i) &= M_1 p_i\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_i,\sigma_i^2}(y_1) \\ \mathcal{F}_k(i) &= M_k \sum_j \mathcal{F}_{k-1}(j) P_{ji}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_i,\sigma_i^2}(y_k) \end{aligned} \tag{A.2}$$

and the *Backward variable*

$$\mathcal{B}_k(i) = \frac{f\left(\boldsymbol{y}_{k+1}^T \mid X_k = i ; \boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_{k+1}^T \mid \boldsymbol{y}_1^k ; \boldsymbol{\theta}\right)} \tag{A.3}$$

which is computed with a backward recurrence

$$\begin{aligned} \mathcal{B}_T(i) &= 1 \\ \mathcal{B}_k(i) &= M_{k+1} \sum_j \mathcal{B}_{k+1}(j)\, P_{ij}(\boldsymbol{\theta}_X)\, \mathcal{G}_{m_j,\sigma_j^2}(y_{k+1}) \end{aligned} \tag{A.4}$$

The M_k quantities, named the *normalizing coefficients*, are given by

$$\begin{aligned} M_1 &= \left\{ \sum_i p_i(\boldsymbol{\theta}_X)\, \mathcal{G}_{m_i,\sigma_i^2}(y_1) \right\}^{-1} \\ M_k &= \left\{ \sum_i \sum_j \mathcal{F}_{k-1}(j)\, P_{ji}(\boldsymbol{\theta}_X)\, \mathcal{G}_{m_i,\sigma_i^2}(y_1) \right\}^{-1} \end{aligned} \tag{A.5}$$

For notation ease, when we have referred to the normalized Forward - Backward algorithm, we have always omitted the *normalized* attribute. However, we must consider that usually the algorithm denoted simply as *Forward - Backward algorithm* is a non normalized version of the above equations (A.1-A.4). The non normalized Forward and Backward variables are respectively given by

$$\tilde{\mathcal{F}}_k(i) = P(X_k = i \,|\, y_1 \ldots y_k, \boldsymbol{\theta}) \tag{A.6}$$

$$\tilde{\mathcal{B}}_k(i) = f(y_{k+1} \ldots y_T \,|\, X_k = i\,;\, \boldsymbol{\theta}) \tag{A.7}$$

They are computed with non normalized recurrence equation which differs from the normalized ones by the absence of the normalizing coefficients M_k. A detailed exposition of the non normalized Forward - Backward algorithm may be found in the work of [Rabiner and Juang, 1986].

The lack of normalizing coefficients cause the algorithm to be numerically instable by underflow. As discussed in [Levinson *et al.*, 1983], such an instability is due to the fact that the non normalized Forward (A.6) and Backward (A.6) variables tend to zero exponentially as the number of iteration increases.

The normalized recurrences, as well as all the expressions used to evaluate the likelihood function or the quantities occurring in its maximization, are a well known issue [Levinson *et al.*, 1983], [Devijver and Dekessel, 1988]. However, for the sake of clarity, we have developed their proofs in the following.

A.2 Forward and Backward Recurrences

We now prove that the Forward and the Backward variables can be respectively computed by a forward recurrence and a backward recurrence, and that these recurrences are the ones given in (A.2) and (A.4).

By subsequently applying the Bayes rule to the Forward variable (A.1)

$$
\begin{aligned}
P\left(X_k = i \mid \boldsymbol{y}_1^k ; \boldsymbol{\theta}\right) &= \frac{1}{f\left(\boldsymbol{y}_1^k ; \boldsymbol{\theta}\right)} \sum_{\boldsymbol{x}_1^k \mid x_k = i} f\left(\boldsymbol{y}_1^k \mid \boldsymbol{x}_1^k ; \boldsymbol{\theta}_{Y \mid X}\right) P\left(\boldsymbol{x}_1^k ; \boldsymbol{\theta}_X\right) \\
&= \sum_{\boldsymbol{x}_1^k \mid x_k = i} \frac{f\left(\boldsymbol{y}_1^{k-1} \mid \boldsymbol{x}_1^{k-1} ; \boldsymbol{\theta}_{Y \mid X}\right)}{f\left(\boldsymbol{y}_1^k ; \boldsymbol{\theta}\right)} f\left(y_k \mid x_k ; \boldsymbol{\theta}_{Y \mid X}\right) P\left(x_k \mid x_{k-1} ; \boldsymbol{\theta}_X\right) P\left(\boldsymbol{x}_1^{k-1} ; \boldsymbol{\theta}_X\right) \\
&= \sum_{\boldsymbol{x}_1^k \mid x_k = i} \frac{f\left(\boldsymbol{y}_1^{k-1}, \boldsymbol{x}_1^{k-1} ; \boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_1^k ; \boldsymbol{\theta}\right)} f\left(\boldsymbol{y}_1^{k-1}, \boldsymbol{x}_1^{k-1} ; \boldsymbol{\theta}\right) P\left(x_k \mid x_{k-1} ; \boldsymbol{\theta}_X\right) f\left(y_k \mid x_k ; \boldsymbol{\theta}_{Y \mid X}\right)
\end{aligned}
$$

we obtain

$$
P\left(X_k = i \mid \boldsymbol{y}_1^k ; \boldsymbol{\theta}\right) = \sum_j \frac{P\left(X_{k-1} = j \mid \boldsymbol{y}_1^{k-1} ; \boldsymbol{\theta}\right)}{f\left(y_k \mid \boldsymbol{y}_1^{k-1} ; \boldsymbol{\theta}\right)} P_{ji}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_k, \sigma_i^2}\left(y_k\right) \tag{A.8}
$$

By substituting $\mathcal{F}_{k-1}(j) = P\left(X_{k-1} = j \mid \boldsymbol{y}_1^{k-1} ; \boldsymbol{\theta}\right)$ and $M_k = f\left(y_k \mid \boldsymbol{y}_1^{k-1} ; \boldsymbol{\theta}\right)^{-1}$, the above equation gives (A.2). The equivalence between (A.5) and

$$
M_k = f\left(y_k \mid \boldsymbol{y}_1^{k-1} ; \boldsymbol{\theta}\right)^{-1}
$$

is deduced by identification from (A.8) and (A.2).

The evaluation of $\mathcal{F}_k(i)$, for $i = 1 \ldots N$ and $k = 1 \ldots N$, through the above recurrence, requires the order of $N^2 T$ calculations (we have NT terms $\mathcal{F}_k(i)$ and from (A.2) each of them requires $2N + 1$ calculations). Note that the quantities $\mathcal{G}_{m_j, \sigma_j^2}(y_k)$ and $P_{ij}\left(\boldsymbol{\theta}_X\right)$ can be computed straightforwardly (in the latter case once the parameterization of the Markov chain is chosen).

Similarly, by subsequently applying the Bayes rule to the Backward variable (A.3)

$$\frac{f\left(\boldsymbol{y}_{k+1}^{T} \mid X_k = i\,;\,\boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_{k+1}^{T} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)} = \frac{1}{f\left(\boldsymbol{y}_{k+1}^{T} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)} \sum_{y_k} \sum_{\boldsymbol{x}_{k+1}^{T}} f\left(\boldsymbol{y}_k^{T} \boldsymbol{x}_{k+1}^{T} \mid X_k = i\,;\,\boldsymbol{\theta}\right)$$

$$= \frac{1}{f\left(\boldsymbol{y}_{k+1}^{T} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)} \sum_{\boldsymbol{x}_{k+1}^{T}} f\left(\boldsymbol{y}_{k+1}^{T} \mid \boldsymbol{x}_{k+1}^{T}\,;\,\boldsymbol{\theta}\right) P\left(\boldsymbol{x}_{k+1}^{T} \mid X_k = i\,;\,\boldsymbol{\theta}_X\right)$$

$$= \sum_{\boldsymbol{x}_{k+1}^{T}} \frac{f\left(\boldsymbol{y}_{k+2}^{T} \boldsymbol{x}_{k+2}^{T} \mid x_{k+1}\,;\,\boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_{k+2}^{T} \mid \boldsymbol{y}_1^{k+1}\,;\,\boldsymbol{\theta}\right) f\left(y_{k+1} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)} f\left(y_{k+1} \mid x_{k+1}\,;\,\boldsymbol{\theta}_{Y\mid X}\right) P\left(x_{k+1} \mid X_k = i\,;\,\boldsymbol{\theta}_X\right)$$

we obtain

$$\frac{f\left(\boldsymbol{y}_{k+1}^{T} \mid X_k = i\,;\,\boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_{k+1}^{T} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)} = \sum_{\boldsymbol{x}_{k+1}^{T}} \frac{f\left(\boldsymbol{y}_{k+2}^{T} \mid X_{k+1} = j\,;\,\boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_{k+2}^{T} \mid \boldsymbol{y}_1^{k+1}\,;\,\boldsymbol{\theta}\right) f\left(y_{k+1} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)} P_{ij}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_j,\sigma_j^2}\left(y_{k+1}\right) \tag{A.9}$$

By substituting $\mathcal{B}_{k+2}\left(i\right) = \dfrac{f\left(\boldsymbol{y}_{k+2}^{T} \mid X_{k+1} = j\,;\,\boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}_{k+2}^{T} \mid \boldsymbol{y}_1^{k+1}\,;\,\boldsymbol{\theta}\right)}$ and $M_{k+1} = f\left(y_{k+1} \mid \boldsymbol{y}_1^k\,;\,\boldsymbol{\theta}\right)^{-1}$ in the above equation, we obtain (A.4).

Akin to the Forward variables, the evaluation of the Backward variables $\mathcal{B}_k\left(i\right)$, for $i = 1 \ldots N$ and $k = 1 \ldots N$, is achieved with a computation burden in the order of N^2T calculations.

A.3 Likelihood Evaluation

In Section 3.3, we have mentioned that a straightfoward evaluation of the likelihood function from its expression is computationally infeasible. Indeed, from the expression of the likelihood function (3.1)

$$f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = \sum_{\boldsymbol{x}} p_{x_1}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_{x_1},\sigma_{x_1}^2}\left(y_1\right) \prod_{k=2}^{T} P_{x_{k-1}x_k}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_{x_k},\sigma_{x_k}^2}\left(y_k\right) \tag{A.10}$$

we can observe a computation burden of TN^T calculations (a summation of a T term product, over the N^T values of $\boldsymbol{x}$). Even for small values of N and T this computation cannot be handle (*e.g.,* for $N = 5$ and $T = 100$ there are the order of 10^{72} computations!).

In order to obtain an efficient algorithm for the likelihood evaluation, a subsequen-

tial application of the Bayes rule to the likelihood function gives

$$f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = f\left(y_1\,;\,\boldsymbol{\theta}\right) \prod_{k=2}^{T} f\left(y_k \,|\, \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right) \tag{A.11}$$

Application of Bayes rule on each $f\left(y_k \,|\, \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right)$ yields the expression

$$\begin{aligned} f\left(y_k \,|\, \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right) &= \frac{1}{f\left(\boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right)} \sum_{\boldsymbol{x}_1^k} f\left(\boldsymbol{y}_1^k \,|\, \boldsymbol{x}_1^k\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right) P\left(\boldsymbol{x}_1^k\,;\,\boldsymbol{\theta}_X\right) \\ &= \frac{1}{f\left(\boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right)} \sum_{\boldsymbol{x}_1^k} f\left(\boldsymbol{y}_1^{k-1}\boldsymbol{x}_1^{k-1}\,;\,\boldsymbol{\theta}\right) P\left(x_k \,|\, x_{k-1}\,;\,\boldsymbol{\theta}_X\right) f\left(y_k \,|\, x_k\,;\,\boldsymbol{\theta}_{Y\,|\,X}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} P\left(X_{k-1} = i \,|\, \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right) P_{ij}\left(\boldsymbol{\theta}_X\right) \mathcal{G}_{m_j,\sigma_j^2}\left(y_k\right) \end{aligned} \tag{A.12}$$

The quantities $\mathcal{G}_{m_j,\sigma_j^2}\left(y_k\right)$ and $P_{ij}\left(\boldsymbol{\theta}_X\right)$ can be computed straightforwardly (the latter once the parameterization of the Markov chain is chosen), while $P\left(X_{k-1} = i \,|\, \boldsymbol{y}_1^{k-1}\,;\,\boldsymbol{\theta}\right)$ is the Forward variable $\mathcal{F}_{k-1}\left(i\right)$ (A.1), and is computed by mean of the forward recurrence (A.2).

From (A.11), (A.12) and (A.5), the likelihood function reads

$$f\left(\boldsymbol{y}\,;\,\boldsymbol{\theta}\right) = \prod_{k=1}^{T} M_k^{-1}$$

The evaluation of the likelihood function is then achieved with a computation burden of the order of N^2T calculations ($3N^2$ calculations for each M_k, $3N^2$ calculation for each $\mathcal{F}_k$, with $k = 1 \ldots T$).

A.4 Computation of $P\left(X_k = i \,|\, \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)$ and $P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)$

With arguments akin to the ones used in the previous section, the straighforward computation of $P\left(X_k = i \,|\, \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)$ and $P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right)$ is practically infeasible.

A subsequential application of Bayes rule to $P\left(X_k = i \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)$ yields

$$P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \frac{1}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)} \sum_{\boldsymbol{x}\,|\,x_k=i} f\left(\boldsymbol{y} \,|\, \boldsymbol{x}\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) P\left(\boldsymbol{x}\,;\, \boldsymbol{\theta}^0_X\right)$$

$$= \sum_{\boldsymbol{x}\,|\,x_k=i} \frac{f\left(\boldsymbol{y}_1^k \,|\, \boldsymbol{x}_1^k\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right)}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)} P\left(\boldsymbol{x}_1^k\,;\, \boldsymbol{\theta}^0_X\right) f\left(\boldsymbol{y}_{k+1}^T \,|\, \boldsymbol{x}_{k+1}^T\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) P\left(\boldsymbol{x}_{k+1}^T \,|\, x_k\,;\, \boldsymbol{\theta}^0_X\right)$$

and by a multiplication with $\sum_{y_k} P\left(y_k \,|\, x_k\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) = 1$ we obtain

$$P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \frac{f\left(\boldsymbol{y}_1^k, X_k = i\,;\, \boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)} \sum_{y_k} \sum_{\boldsymbol{x}_{k+1}^T\,|\,x_k=i} f\left(y_k, \boldsymbol{y}_{k+1}^T, \boldsymbol{x}_{k+1}^T \,|\, x_k\,;\, \boldsymbol{\theta}^0\right)$$

$$= f\left(X_k = i \,|\, \boldsymbol{y}_1^k\,;\, \boldsymbol{\theta}^0\right) \frac{f\left(\boldsymbol{y}_{k+1}^T \,|\, x_k\,;\, \boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{y}_{k+1}^T \,|\, \boldsymbol{y}_1^k\,;\, \boldsymbol{\theta}^0\right)} \tag{A.13}$$

From the definition of the Forward variable (A.1) and the Backward variable (A.3), the above equation reads

$$P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \mathcal{F}_k^0\left(i\right) \mathcal{B}_k^0\left(i\right)$$

Similarly, a subsequential application of Bayes rule to $P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)$ yields

$$P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \frac{1}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)} \sum_{\boldsymbol{x}\,|\,\substack{x_k=j \\ x_{k-1}=i}} f\left(\boldsymbol{y} \,|\, \boldsymbol{x}\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) P\left(\boldsymbol{x}\,;\, \boldsymbol{\theta}^0_X\right)$$

$$= \frac{1}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)} \sum_{\boldsymbol{x}\,|\,\substack{x_k=j \\ x_{k-1}=i}} f\left(\boldsymbol{y}_1^{k-1} \,|\, \boldsymbol{x}_1^{k-1}\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) f\left(\boldsymbol{y}_k^T \,|\, \boldsymbol{x}_k^T\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) P\left(\boldsymbol{x}\,;\, \boldsymbol{\theta}^0_X\right)$$

$$= \sum_{\boldsymbol{x}\,|\,\substack{x_k=j \\ x_{k-1}=i}} \frac{f\left(\boldsymbol{y}_1^{k-1}, \boldsymbol{x}_1^{k-1}\,;\, \boldsymbol{\theta}^0\right) f\left(\boldsymbol{y}_{k+1}^T, \boldsymbol{x}_{k+1}^T \,|\, x_k\,;\, \boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{y}\,;\, \boldsymbol{\theta}^0\right)} f\left(y_k \,|\, x_k\,;\, \boldsymbol{\theta}^0_{Y\,|\,X}\right) P\left(x_k \,|\, x_{k-1}\,;\, \boldsymbol{\theta}^0_X\right)$$

and finally

$$P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) = P\left(X_{k-1} = i \,|\, \boldsymbol{y}_1^{k-1}\,;\, \boldsymbol{\theta}^0\right) P_{ij}\left(\boldsymbol{\theta}_X^0\right) \mathcal{G}_{m_j^0, \sigma_j^{2\,0}}\left(y_k\right) \frac{f\left(\boldsymbol{y}_{k+1}^T \,|\, X_k = j\,;\, \boldsymbol{\theta}^0\right)}{f\left(\boldsymbol{y}_k^T \,|\, \boldsymbol{y}_1^{k-1}\,;\, \boldsymbol{\theta}^0\right)}$$

From the definition of the Forward variable (A.1) and the Backward variable (A.3) the above equation reads

$$P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right) = \mathcal{F}_{k-1}^0\left(i\right) P_{ij}\left(\boldsymbol{\theta}_X^0\right) \mathcal{G}_{m_j^0, \sigma_j^{2\,0}}\left(y_k\right) \mathcal{B}_k^0\left(i\right) \tag{A.14}$$

Note that the evaluation of $P\left(X_k = i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)$ or $P\left(X_{k-1} = i, X_k = j \,|\, \boldsymbol{y}, \boldsymbol{\theta}^0\right)$ is achieved with a computation burden in the order of NT calulations (N calculations for T normalizing coefficients and $2N$ calculation for each one of the Forward and Backward variables).

Appendix B

Segmentation Program

B.1 Introduction

In this appendix we briefly describe the programming of the segmentation method that we have studied, and we provide a representation of the program structure.

The program can be considered composed of two main parts, represented in figure B.1 and figure B.2, respectively: the first part estimates the model parameters, while the second one computes the segmented image.

We have programmed the segmentation method mostly with the *Matlab* language. Due to the computational burden, the Forward - Backward algorithm and the algorithm for the computation of the maximum *a posteriori* have been programmed with the *C* language, and then they have been integrated to the *Matlab* program.

B.2 Program for the Parameter Estimation

In the following, we refer to the hidden Markov model based on Pickard random fields (see Chapter 5), which provides the necessary characteristics for the unidimensional aproach to unsupervised image segmentation. In particular, the parameters that we aim to estimate are the parameters $\boldsymbol{\theta}_{Y|X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ of the observable characteristic, and the telegraphic parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ of the hidden process.

The first part of the program computes the maximum likelihood estimates of the model parameters, by means of a fixed point method [Frontini, 1996].

For the estimation of the parameters $\boldsymbol{\theta}_{Y|X} = \{\boldsymbol{m}, \boldsymbol{\sigma}^2\}$ we refer to the Baum-Welch re-estimation formulas (see Section 3.5 and Section 6.2.1): they constitute a well studied classical method for the re-estimation of such parameters.

For the estimation of the telegraphic parameters $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ we do not have a satisfactory re-restimation method such as the Baum-Welch re-estimation formulas. In order to find the method that is the most adapted to our problem we consider three different re-estimation methods:

- The first re-estimation method is provided by the telegraphic EM algorithm that has been defined in Chapter 4.

- The second re-estimation method is a mixed TEM - gradient descent technique: we consider the re-estimation formulas of the TEM algorithm until the values of the test variables stabilize, then, we apply the gradient descent method.

  The gradient descent method that we have used is a very basic one, with a non optimized descent step. The descent step is initialize with an empirical value and then reduced every time that the likelihood function is not increased or the parameter constraints are not fulfilled.

  Note that the gradient of the likelihood is provided by Property 3.6.

- The third re-estimation method is the basic gradient descent technique that we have described above. The initialization of the method is the same as the TEM algorithm.

The first step of the parameter estimation algorithm is the initialization of the parameters: $\boldsymbol{\theta}^0 = \left\{\boldsymbol{\theta}^0_{Y\,|\,X}, \boldsymbol{\theta}^0_X\right\}$. Then, one enters the loop where the Forward and Backward variables $\mathcal{F}^0$, $\mathcal{B}^0$ are computed in order to evaluate the parameter estimates by means of the re-estimation formulas: $\widehat{\boldsymbol{\theta}} = g\left(\mathcal{F}^0, \mathcal{B}^0, \boldsymbol{\theta}^0\right)$ (see Chapter 6).

The decision of whether or not exit the loop is taken according to the values of some test variables, which control if the fixed point is attained. These variable are:

$$\frac{|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0|}{|\widehat{\boldsymbol{\theta}}|} \tag{B.1}$$

$$\frac{|f(\boldsymbol{y}\,;\,\widehat{\boldsymbol{\theta}}) - f(\boldsymbol{y}\,;\,\boldsymbol{\theta}^0)|}{|f(\boldsymbol{y}\,;\,\widehat{\boldsymbol{\theta}})|} \tag{B.2}$$

$$\frac{|\nabla f(\boldsymbol{y}\,;\,\widehat{\boldsymbol{\theta}}) - \nabla f(\boldsymbol{y}\,;\,\boldsymbol{\theta}^0)|}{|\nabla f(\boldsymbol{y}\,;\,\widehat{\boldsymbol{\theta}})|} \tag{B.3}$$

In our program we exit the loop if all the above variable are less than 10^{-4}, 10^{-4} and 10^{-6}, respectively. Once the iterative procedure has terminated, we obtain the parameter estimate $\widehat{\boldsymbol{\theta}}$.

Figure B.1 provides a graphical representation of the parameter estimation program.

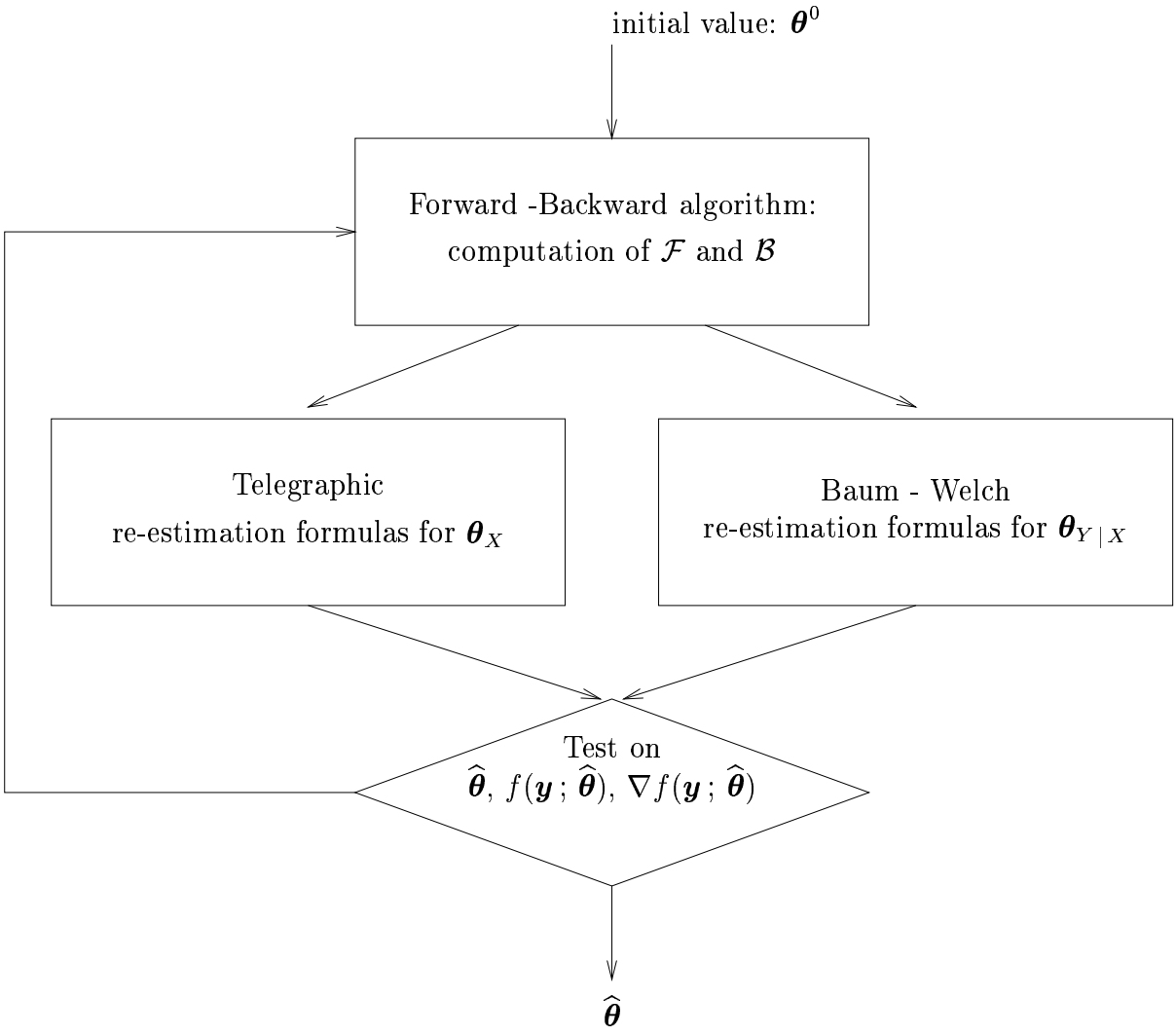

Figure B.1: Structure of the Program for the Parameter Estimation

B.3 Program for the Segmented Image Computation

The second part of the program computes the segmented image by a maximum *a posteriori* technique on the marginal *a posteriori* probability.

Once the parameters are estimated, by means of the computation of the Forward and Backward variables $\mathcal{F}$, $\mathcal{B}$ we have access to the marginal *a posteriori* distribution of the observable process: $P(X_{r,c}\,|\,\boldsymbol{y}\,;\,\widehat{\boldsymbol{\theta}})$. The segmented image $\boldsymbol{x}_{\text{seg}} = \left\{x_{\text{seg}_{r,c}}\right\}_{(r,c)\in\Lambda}$

is then obtained by means of the maximization of such a distribution (see Section 7.3):

$$x_{\mathrm{seg}_{r,c}} = \underset{i \in \{1,\dots,N\}}{\arg\max}\, P(X_{r,c} \,|\, \boldsymbol{y}\,;\, \widehat{\boldsymbol{\theta}})\,;\, (r,c) \in \Lambda$$

Figure B.1 provides a graphical representation of the program for the maximum *a posteriori* (MAP) computation.

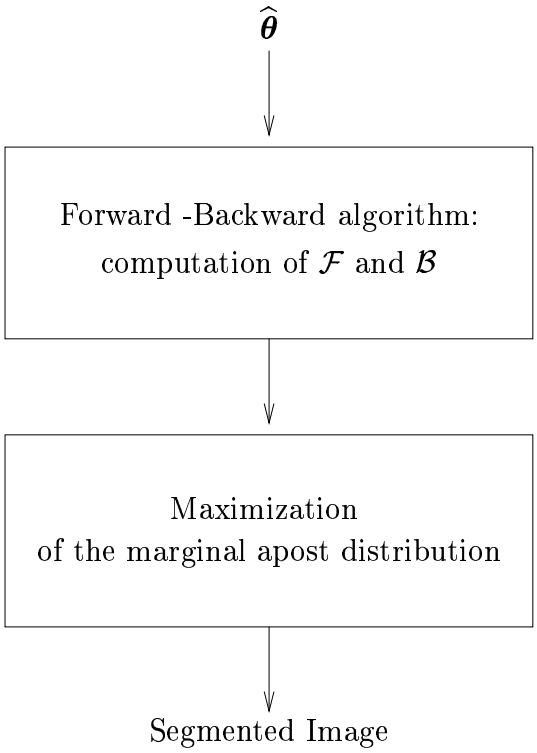

Figure B.2: Structure of the Program for the Segmented Image Computation

Bibliography

[Baum *et al.*, 1970] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[Besag *et al.*, 1991] Julian E. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of Institute of Statistical Mathematics*, 43(1):1–59, 1991.

[Bittanti, 1993] Sergio Bittanti. *Teoria della Predizione e del Filtraggio.* Pitagora Editrice, Bologna, Italy, 1993.

[Brémaud, 1997] Pierre Brémaud. *Markov Chains, Gibbs Fields, Stochastic Network and Monte Carlo Simulation.* preprint, 1997.

[Campillo and Le Gland, 1989] Fabien Campillo and François Le Gland. MLE for partially observed diffusions: direct maximization vs. EM algorithm. *Stochastic Processes and their Application*, 33:245–274, 1989.

[Champagnat *et al.*, 1998] Frédéric Champagnat, Jérôme Idier, and Yves Goussard. Stationary Markov random fields on a rectangular finite lattice. Technical report, IGB-LSS, IEEE Trans. Inf. Theory, 1998.

[Dacunha-Castelle and Duflo, 1982] Didier Dacunha-Castelle and Marie Duflo. *Probabilités et statistiques, 1. Problèmes à temps fixe.* Masson, Paris, 1982.

[Demoment, 1996] Guy Demoment. *Traitement Statistique du Signal, notes du cours. DEA Automatique et Traitement du Signal*, Université de Paris Sud - Orsay, 1996.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38, 1977.

[Devijver and Dekessel, 1988] P. A. Devijver and M. Dekessel. Champs aléatoires de Pickard et modélisation d'images digitales. *Traitement du Signal*, 5(5):131–150, 1988.

[Diebolt and Robert, 1994] Jean Diebolt and Christian P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B*, 56(2):363–375, 1994.

[Frontini, 1996] Marco Frontini. *Calcolo Numerico, appunti del corso. Politecnico di Milano*, Milan, Italy, 1996.

[Geman and Geman, 1984] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6(6):721–741, November 1984.

[Godfrey *et al.*, 1980] R. Godfrey, F. Muir, and F. Rocca. Modeling seismic impedance with Markov chains. *Geophysics*, 45(9):1351–1372, September 1980.

[Goussard *et al.*, 1997] Yves Goussard, Jérôme Idier, and Alain De Cesare. Unsupervised image segmentation using a telegraph parameterization of Pickard random fields. In *Proc. IEEE ICASSP*, pages 2777–2780, Munich, Germany, April 1997.

[Hero and Fessler, 1985] Alfred O. Hero and Jeffrey A. Fessler. Asymptotic convergence properties of EM-type algorithms. Preprints 85-T-21, Dept. of Electrical Engineering and Computer Science, University of Michigan, 1985.

[Idier and Goussard, 1995] Jérôme Idier and Yves Goussard. Formules de réestimation pour un modèle de chaîne de Markov cachée stationnaire réversible. In *Actes du 15^{e} colloque GRETSI*, pages 177–180, Juan-les-Pins, France, September 1995.

[Leroux and Puterman, 1992] Brian G. Leroux and Martin L. Puterman. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48:545–558, June 1992.

[Levinson *et al.*, 1983] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, April 1983.

[Liporace, 1982] Luis A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Trans. Inf. Theory*, IT-28:729–734, September 1982.

[Meilijson, 1989] I. Meilijson. A fast improvement of the EM algorithm on its own terms. *J. R. Statist. Soc. B*, 51:127–138, 1989.

[Nádas, 1983] Arthur Nádas. Hidden Markov chains, the forward-backward algorithm, and initial statistics. *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-31(2):504–506, April 1983.

[Pickard, 1980] David K. Pickard. Unilateral Markov fields. *Adv. Appl. Prob.*, 12:655–671, 1980.

[Rabiner and Juang, 1986] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP magazine*, pages 4–16, 1986.

[Redner and Walker, 1984] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, 26(2):195–239, April 1984.

[Robert, 1992] Christian Robert. *L'analyse statistique Bayésienne.* Economica, 1992.

[Woess, 1996] Wolfang Woess. *Processi Stocastici I, dispense del corso. Dipartimento di Matematica, Università Statale di Milano*, Milan, Italy, 1996.